GA-NAR: GEOMETRY-ADAPTIVE NONLINEARITY ALLOCATION AND REPRESENTATION FOR COMPLEX PHYSICAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many complex physical systems exhibit a recurring dominant nonlinear-response structure whose local intensity and dominant/complement allocation vary across operating conditions, while standard supervised surrogates typically apply a fixed pattern of nonlinear computation. We introduce GA-NAR (Geometry-Adaptive Nonlinearity Allocation and Representation), a backbone-compatible framework that adapts neural nonlinear computation to the structure of the physical parameter-to-solution map. GA-NAR comprises two coupled components: Nonlinear Response Geometry Decomposition (NRGD), which identifies a recurring Dominant Nonlinear-Response Subspace together with state-dependent response intensity and subspace allocation, and Geometry-Adaptive Nonlinear Representation (GANR), which uses this information to organize nonlinear computation across subspaces and network depth. Independent experiments on NRGD across five systems verify reliable recovery of the proposed response descriptors, while separate no-physics and shared-physics factorial experiments show that GANR achieves the lowest NRMSE on four of five auxiliary systems under both conditions. The complete GA-NAR framework reduces standardized test MSE in all 18 activation–system comparisons, with median reductions of 11.5%, 2.1%, and 2.8% on IEEE-118, Kuramoto-64, and nonlinear heat FEM, respectively, alongside larger activation-specific gains in several cases. Physical residuals also decrease for most tested activations, showing that the predictive gains generally extend to physical consistency. We will release codes at https://anonymous.4open.science/r/GA-NAR.

1 INTRODUCTION

Complex physical systems can exhibit sharply varying parameter-to-solution responses across operating conditions. Neural operators such as DeepONet and FNO amortize repeated solves (Lu et al., 2021; Li et al., 2021), supported by general operator-learning theory and benchmarks (Kovachki et al., 2023; Takamoto et al., 2022); recent extensions use Laplace representations, pretraining, and latent dynamics (Cao et al., 2024; Hao et al., 2024a; Kontolati et al., 2024). Architectures also encode locality, domain geometry, conservation, long-range interactions, equivariance, or derivative structure (Li et al., 2020; 2023b; Cheng et al., 2024; Liu-Schiaffini et al., 2024; Liu et al., 2024; Lingsch et al., 2024; Shumaylov et al., 2025; Liu et al., 2026). These advances improve scientific surrogate learning but do not determine how nonlinear computation should respond to a recurring dominant nonlinear-response structure together with state-dependent response intensity and dominant/complement allocation.

Pointwise accuracy does not resolve this question. Derivative-informed learning shows that similar value fits can have different Jacobian or sensitivity fidelity (Czarnecki et al., 2017; O'Leary-Roseberry et al., 2022; 2024; Qiu et al., 2024; Behroozi et al., 2025), with consequences for optimization and inverse problems (Cho et al., 2025; Cheng et al., 2025; Gonzalez-Sieiro et al., 2026). Physics-informed learning constrains governing residuals (Raissi et al., 2019), but low residual and accurate recovery need not coincide (Krishnapriyan et al., 2021; Wang et al., 2022); certification, benchmarking, and uncertainty studies likewise distinguish predictive, physical, and response fidelity (Eiras et al., 2024; Hao et al., 2024b; Magnani et al., 2025). We therefore ask whether a reusable backbone can adapt its nonlinear computation to finite-scale physical response structure.

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.

GA-NAR separates system-side response measurement from model-side nonlinear allocation. NRGD extracts a dominant response subspace, local response intensity, and dominant/complement allocation from the physical map; GANR uses these descriptors to modulate nonlinear preactivations across depth while leaving the activation family unchanged. Unlike generic conditioning through FiLM (Perez et al., 2018), adapters, or low-rank adaptation (Houlsby et al., 2019; Hu et al., 2022; Zhang et al., 2023; Wang et al., 2024), the control variables originate from physical response structure. We evaluate the same residual-backbone interface (He et al., 2016) with ReLU, GELU, SiLU, Mish, Softplus, and SwiGLU (Hendrycks & Gimpel, 2016; Ramachandran et al., 2018; Misra, 2020; Shazeer, 2020), treating activation family as a controlled factor (Apicella et al., 2021).

Our contributions are:

- **NRGD.** We propose Nonlinear Response Geometry Decomposition (NRGD) to expose the nonlinear-response structure that a neural surrogate must represent. From finite-scale Jacobian variation, NRGD identifies a basis-invariant, globally recurring Dominant Nonlinear-Response Subspace (DRS) together with state-dependent Local Nonlinear-Response Intensity and Local Subspace Response Allocation, providing system-side guidance for neural nonlinear representation.
- **GANR.** We propose Geometry-Adaptive Nonlinear Representation (GANR) to improve a neural backbone's ability to represent physical-system nonlinearity. GANR uses the response information identified by NRGD through a dual-subspace architecture that preserves dominant and complementary response channels, together with layer-wise geometry modulation that adapts their contribution and modulation strength across network depth without changing the backbone activation family.
- **GA-NAR.** We propose GA-NAR as a backbone-compatible mechanism for adapting nonlinear computation to physical response structure. The complete method reduces standardized test MSE in all 18 activation–system comparisons, with median reductions of 11.5%, 2.1%, and 2.8% on IEEE-118, Kuramoto-64, and nonlinear heat FEM, respectively, alongside larger activation-specific gains; physical residuals also decrease for most tested activations, showing favorable physical consistency across most tested settings.

2 Related Work

Scientific surrogate and operator learning. Neural operators span DeepONet/FNO, graph and transformer architectures, pretraining, and geometry- or physics-structured designs (Lu et al., 2021; Li et al., 2021; 2020; 2023a; Hao et al., 2024a; Chen et al., 2024; Li et al., 2023b; Cheng et al., 2024; Liu-Schiaffini et al., 2024; Xiao et al., 2024; Lingsch et al., 2024; Liu et al., 2024; Singh et al., 2024; Gopakumar et al., 2026).

Physics and derivative informed learning. Physics- and derivative-informed methods constrain equations or supervise derivatives, Jacobian actions, and sensitivities (Raissi et al., 2019; Cho et al., 2024; Hao et al., 2024b; Czarnecki et al., 2017; O'Leary-Roseberry et al., 2022; 2024; Qiu et al., 2024; Behroozi et al., 2025; Cho et al., 2025; Cheng et al., 2025; Gonzalez-Sieiro et al., 2026). GA-NAR instead conditions nonlinear computation on finite-scale variation of the physical parameter-to-solution Jacobian.

Subspaces and adaptive representation. Active-subspace and derivative/Hessian methods identify influential directions or reduced parameterizations (Constantine, 2015; Chen & Ghattas, 2019; Luo et al., 2026); NRGD instead constructs its Dominant Nonlinear-Response Subspace from finite-scale variation of the solution-map Jacobian, rather than from first-order sensitivity such as $J_G^\top J_G$, and retains the complement for local response allocation. Domain/symmetry conditioning (Lee et al., 2026; Shumaylov et al., 2025) and adaptive activations, FiLM, adapters, and low-rank adaptation (Jagtap et al., 2020a;b; Apicella et al., 2021; Perez et al., 2018; Houlsby et al., 2019; Hu et al., 2022; Zhang et al., 2023; Wang et al., 2024) do not provide this system-derived dual-subspace allocation; GANR uses NRGD intensity/allocation and learns its layer-wise realization. Multi-criterion evaluation follows prior benchmarking, certification, hybrid, uncertainty, latent-dynamics, and derivative-reconstruction studies (Takamoto et al., 2022; Eiras et al., 2024; Rudikov et al., 2024; Magnani et al., 2025; Kontolati et al., 2024; Liu et al., 2026).

3 PROBLEM FORMULATION

3.1 IMPLICIT PHYSICAL SURROGATE LEARNING

We use s to index a physical system. For system s, $q \in \mathbb{R}^{d_s}$ denotes the physical parameter vector with input dimension d_s, and $z \in \mathbb{R}^{m_s}$ denotes the physical state with state dimension m_s. The symbol $\mathcal{F}_s$ denotes the governing residual map of that system. Training-only statistics provide parameter and state means $\mu_{q,s}$ and $\mu_{z,s}$ together with diagonal coordinate-scale matrices $D_{q,s}$ and $D_{z,s}$. We use $x \in \mathbb{R}^{d_s}$ and $y \in \mathbb{R}^{m_s}$ for the corresponding standardized parameter and state coordinates, and G_s for the protocol-selected local parameter-to-solution map in standardized coordinates. These objects are related by

$$\mathcal{F}_s(z, q) = 0, \qquad x = D_{q,s}^{-1}(q - \mu_{q,s}), \qquad y = D_{z,s}^{-1}(z - \mu_{z,s}), \qquad y = G_s(x). \tag{1}$$

The map G_s is understood locally on protocol-selected regular solution branches rather than as a globally single-valued map over the entire parameter domain.

At a regular operating point, $\bar{F}_s$ denotes the governing residual expressed in standardized coordinates. We write $J_{G_s}(x) \in \mathbb{R}^{m_s \times d_s}$ for the Jacobian of G_s at standardized input x, so J_{G_s} represents local parameter-to-state sensitivity. The matrices $\bar{F}_{s,y}$ and $\bar{F}_{s,x}$ denote the Jacobians of $\bar{F}_s$ with respect to the standardized state y and standardized input x, respectively. The implicit-function relation gives

$$J_{G_s}(x) = \frac{\partial G_s(x)}{\partial x} = -\bar{F}_{s,y}^{-1}\bar{F}_{s,x}. \tag{2}$$

3.2 THE REPRESENTATION GAP

A conventional supervised surrogate is written as $\widehat{G}_\theta$, where θ denotes its trainable parameters. We denote the supervised-data distribution for system s by Π_s^{data}, the prediction loss by $\mathcal{L}_{\mathrm{pred}}$, the expectation operator by $\mathbb{E}$, and the Euclidean norm by $\|\cdot\|_2$. The notation $x \sim \Pi_s^{\mathrm{data}}$ means that the standardized input x is drawn from this supervised-data distribution. Its primary training objective is value accuracy,

$$\mathcal{L}_{\mathrm{pred}}(\theta) = \mathbb{E}_{x \sim \Pi_s^{\mathrm{data}}}\left[\frac{1}{m_s}\|\widehat{G}_\theta(x) - G_s(x)\|_2^2\right]. \tag{3}$$

Value accuracy alone does not determine how the learned map changes locally. We use $n \in \mathbb{R}^{d_s}$ for a unit perturbation direction and $h > 0$ for a finite perturbation scale in standardized input coordinates. The symbol $\Delta_{G_s}^h(x, n)$ denotes the centered finite-scale variation of the physical Jacobian, while $r_{G_s}^h(x, n)$ denotes its squared directional response magnitude. The Frobenius norm is written as $\|\cdot\|_F$. These quantities are defined by

$$\Delta_{G_s}^h(x, n) = \frac{J_{G_s}(x + hn) - J_{G_s}(x - hn)}{2h}, \qquad r_{G_s}^h(x, n) = \|\Delta_{G_s}^h(x, n)\|_F^2. \tag{4}$$

Throughout the paper, the recurring response subspace and the state-dependent intensity/allocation descriptors are defined relative to the training-standardized input metric and the protocol-selected finite scale; they are not intended as intrinsic coordinate-free curvature quantities.

A learned surrogate can satisfy the value objective in Equation 3 without reproducing the physical response in Equation 4; Proposition 1 in Appendix D formalizes this non-identifiability. GA-NAR therefore conditions nonlinear computation on this recurring-subspace/local-descriptor structure rather than adding a mandatory response-matching loss.

4 GA-NAR METHOD

4.1 OVERVIEW

GA-NAR addresses one problem in two inseparable stages. NRGD identifies how nonlinear response is organized in the physical parameter-to-solution map; GANR determines how that structure should be realized as nonlinear computation inside the backbone. The physical solver and physical Jacobians

are used only to construct training targets and are not queried at inference. Figure 1 summarizes this system-to-network pathway.

NRGD does more than estimate a scalar measure of nonlinearity. From finite-scale Jacobian variation it identifies a basis-invariant **Dominant Nonlinear-Response Subspace (DRS)**, capturing the recurring parameter directions along which nonlinear response is organized. The DRS projector is global over the training geometry population; the operating-point dependence is carried by the local intensity and allocation descriptors. It then measures **Local Nonlinear-Response Intensity** and **Local Subspace Response Allocation**, which respectively quantify how strongly the map responds at the current operating point and how that response is divided between the DRS and its complementary subspace. Retaining the complement is essential: the DRS summarizes recurrent structure without turning geometry extraction into lossy dimension reduction.

GANR is designed as the matching model-side realization of these quantities. The DRS projector defines a **dual-subspace** representation with separate dominant and complementary channels; NRGD intensity and local subspace allocation supervise the amount and composition of the modulation; and a learned **layer-wise geometry modulation** distributes this response-aware signal across network depth, where no physical target exists. The resulting signal is RMS-matched to each layer and injected at the preactivation level, so GA-NAR reallocates nonlinear computation without replacing the backbone activation. For SwiGLU, only the gate branch is modulated. Thus NRGD and GANR are linked descriptor by descriptor: physical response geometry is first decomposed into orientation, local strength, and subspace allocation, then translated into representation channels, modulation magnitude/composition, and depth-wise nonlinear computation. Construction and resolution criteria are detailed in Appendix J, and the exact backbone interface is given in Appendix A.3.

For compact notation, we suppress the system index s below. The standardized input and output dimensions are d and m, and G denotes the physical parameter-to-solution map. The finite-scale Jacobian variation and response score are Δ_G^h and r_G^h. Let R be the selected response rank and $P_R \in \mathbb{R}^{d\times d}$ the rank-R DRS projector. NRGD produces the local intensity target $\rho_G^\star(x)$ and the two-component dominant/complement allocation $\boldsymbol{\pi}_G^\star(x)$. GANR predicts their neural counterparts and distributes the resulting modulation across L nonlinear blocks.

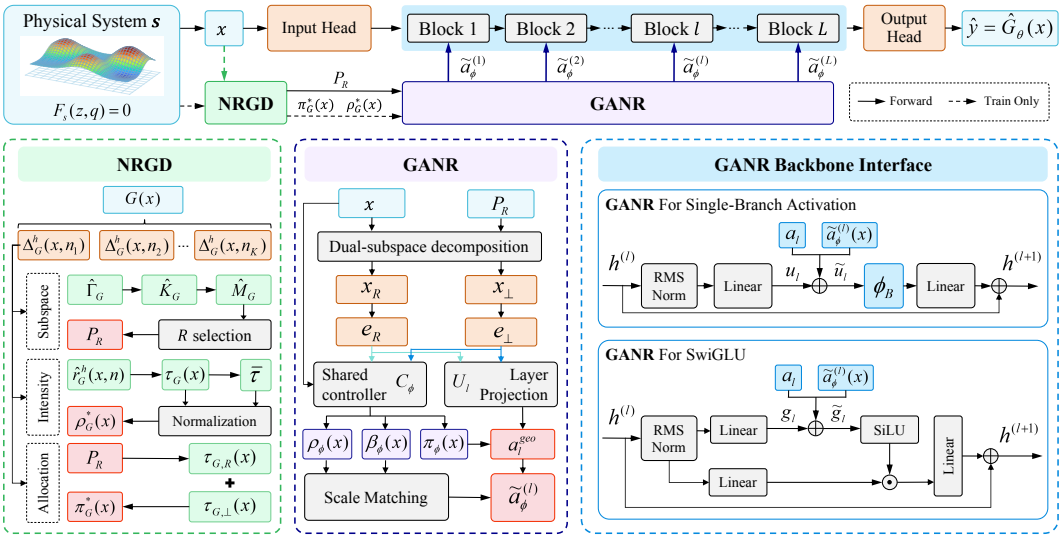


Figure 1: GA-NAR architecture. NRGD extracts physical response geometry, and GANR converts it into layer-wise nonlinear modulation.

4.2 NRGD: Nonlinear Response Geometry Decomposition

NRGD starts from the finite-scale response defined in Section 3. Its central objective is to recover the recurring directional organization of nonlinear response without collapsing the remaining response

into discarded dimensions. It therefore constructs the Dominant Nonlinear-Response Subspace and couples it with local response intensity and Local Subspace Response Allocation. Together these quantities describe the orientation, strength, and dominant/complement composition of physical nonlinearity while avoiding explicit dense Hessians or Jacobians.

Dominant Nonlinear-Response Subspace. NRGD avoids explicitly forming dense Jacobians or Hessians. Let K_H be the number of frozen Hutchinson probes and let $v_k \in \{-1, +1\}^d$ be the kth Rademacher probe, with probe index $k = 1, \ldots, K_H$. Let N_G denote the number of training-only geometry centers, let x_i denote the ith center, and let K_S be the number of frozen unit-sphere directions assigned to each center. The center index i runs from 1 to N_G, and the direction index j runs from 1 to K_S; $s_{ij} \in \mathbb{R}^d$ denotes direction j at center i. The scale $h_\pi^\star > 0$ is the selected finite scale used for subspace allocation. We use the superscript $\top$ to denote vector or matrix transpose. Following the Hutchinson trace-estimation principle (Hutchinson, 1990), a hat indicates an empirical quantity estimated from the frozen geometry samples. We denote the stochastic directional-response estimate by $\widehat{r}_G^h$, the global response moment by $\widehat{S}_G$, and its trace-normalized form by $\widehat{\Gamma}_G$; $\mathrm{tr}(\cdot)$ denotes the matrix trace. These quantities are

$$\begin{aligned} \widehat{r}_G^h(x,n) &= \frac{1}{K_H}\sum_{k=1}^{K_H} \|\Delta_G^h(x,n)v_k\|_2^2, \qquad \widehat{\Gamma}_G = \frac{\widehat{S}_G}{\mathrm{tr}(\widehat{S}_G)}, \\ \widehat{S}_G &= \frac{1}{N_G K_S}\sum_{i=1}^{N_G}\sum_{j=1}^{K_S} \widehat{r}_G^{h_\pi^\star}(x_i, s_{ij}) s_{ij} s_{ij}^\top. \end{aligned} \tag{5}$$

The smooth-limit spherical identity motivates a quadratic-response backprojection: instead of assembling a Hessian, NRGD reconstructs a normalized response operator from response-weighted directional second moments. Let I_d denote the $d \times d$ identity matrix, $\widehat{K}_G$ the backprojected finite-scale response operator, and Π_{Δ_+} the operator that projects its eigenvalues onto the nonnegative unit simplex Δ_+. The projected positive-semidefinite unit-trace operator is denoted by $\widehat{M}_G$. For this operator, $\lambda_i(\widehat{M}_G)$ denotes its ith largest eigenvalue, and $U_{1:R} \in \mathbb{R}^{d\times R}$ collects the corresponding first R orthonormal eigenvectors. In the rank-selection expression below, r is a candidate rank index. The response rank R is chosen as the smallest candidate rank that captures at least 90% of the projected spectral mass, after which the DRS projector P_R is constructed as

$$\begin{aligned} \widehat{K}_G &= \frac{(d+2)\widehat{\Gamma}_G - I_d}{2}, \qquad \widehat{M}_G = \Pi_{\Delta_+}(\widehat{K}_G), \\ R &= \min\left\{ r : \sum_{i=1}^{r} \lambda_i(\widehat{M}_G) \geq 0.9 \right\}, \qquad P_R = U_{1:R}U_{1:R}^\top. \end{aligned} \tag{6}$$

The projector P_R, rather than a particular eigenvector basis, is the object used downstream. This makes the descriptor insensitive to rotations within the selected eigenspace. The complementary projector $I_d - P_R$ is retained explicitly: NRGD is therefore not an active-subspace truncation, but a two-channel decomposition in which the dominant and complementary responses can receive different nonlinear computation.

Local Nonlinear-Response Intensity. The global projector identifies dominant response orientation but does not measure how strongly the physical map is nonlinear at a particular operating point. Let $h_\rho^\star > 0$ denote the separately calibrated finite scale used for intensity estimation, and let s_j denote the jth frozen unit-sphere direction at the current input. We write $\tau_G(x)$ for the local total-response estimate, $\bar{\tau}$ for its mean over the N_G training geometry centers, and $\rho_G^\star(x)$ for the resulting bounded relative intensity target. These quantities are

$$\tau_G(x) = \frac{d}{K_S}\sum_{j=1}^{K_S} \widehat{r}_G^{h_\rho^\star}(x, s_j), \qquad \bar{\tau} = \frac{1}{N_G}\sum_{i=1}^{N_G}\tau_G(x_i), \qquad \rho_G^\star(x) = \frac{\tau_G(x)}{\tau_G(x) + \bar{\tau}}. \tag{7}$$

The selected intensity scale $h_\rho^\star$ is allowed to differ from the allocation scale $h_\pi^\star$. This decouples the scale at which local response magnitude is most reliably resolved from the scale used to estimate

directional allocation, rather than forcing one perturbation radius to serve two different statistical tasks. The bounded quantity $\rho_G^\star$ is interpreted as relative local nonlinear-response intensity with respect to the training geometry population. It is not interpreted as physical energy or as a ground-truth measure of neural capacity.

Local Subspace Response Allocation. NRGD next measures how the local response is distributed between the DRS and its orthogonal complement. Let K_R and $K_\perp$ denote the numbers of frozen probe directions used in the dominant and complementary subspaces. The unit vector $n_{R,j}$ denotes the jth probe in $\mathrm{range}(P_R)$, while $n_{\perp,j}$ denotes the jth probe in $\mathrm{range}(I_d - P_R)$. We write $\tau_{G,R}(x)$ and $\tau_{G,\perp}(x)$ for the corresponding dimension-corrected local response totals, and $\boldsymbol{\pi}_G^\star(x) = (\pi_R^\star(x), \pi_\perp^\star(x))$ for the normalized two-component allocation target. They are defined by

$$\tau_{G,R}(x) = \frac{R}{K_R}\sum_{j=1}^{K_R} \widehat{r}_G^{h_\pi^\star}(x, n_{R,j}), \qquad \tau_{G,\perp}(x) = \frac{d-R}{K_\perp}\sum_{j=1}^{K_\perp} \widehat{r}_G^{h_\pi^\star}(x, n_{\perp,j}),$$
$$\boldsymbol{\pi}_G^\star(x) = \frac{(\tau_{G,R}(x), \tau_{G,\perp}(x))}{\tau_{G,R}(x) + \tau_{G,\perp}(x)}. \tag{8}$$

The dimension factors compensate for different subspace dimensions in the isotropic-probe limit, preventing the larger subspace from receiving more response mass merely because it contains more probe directions. NRGD therefore separates three questions—dominant orientation, local response strength, and local dominant/complement composition—without forcing any one of them to stand in for the others. The geometry descriptors are used only when the prespecified resolution tests in Appendix J are satisfied; otherwise the corresponding quantity is treated as unresolved.

4.3 GANR: Geometry-Adaptive Nonlinear Representation

GANR solves the complementary translation problem: NRGD describes nonlinear response in physical parameter space, whereas a backbone must realize that information in layer-specific preactivation coordinates. GANR therefore mirrors NRGD with a dual-subspace representation and adds layer-wise geometry modulation: dominant/complement response content is preserved in separate channels, local modulation magnitude and composition are tied to NRGD targets, and the network learns at which depths that response-aware computation should be expressed. Depth allocation is learned from the predictive task because neural depth is not a physical coordinate.

Dual-subspace geometry encoding. The standardized input is decomposed by the DRS projector and its complement. We denote the dominant and complementary input components by x_R and $x_\perp$, respectively. Two independent trainable encoders, denoted by E_R and $E_\perp$, map these components to geometry codes e_R and $e_\perp$. The encoding step is

$$x_R = P_R x, \qquad x_\perp = (I_d - P_R)x, \qquad e_R = E_R(x_R), \qquad e_\perp = E_\perp(x_\perp). \tag{9}$$

The encoder pair preserves the complementary component rather than treating the DRS as a lossy dimensionality reduction. Because the decomposition depends on the projector, the construction is invariant to a rotation of the chosen basis within either subspace.

Geometry control variables. Let ϕ denote the trainable GANR parameters. The controller map C_ϕ combines the standardized input and the two geometry codes into a shared controller representation g_ϕ. We use $\rho_\phi(x) \in (0,1)$ for the predicted response intensity, $\boldsymbol{\pi}_\phi(x) = (\pi_{\phi,R}(x), \pi_{\phi,\perp}(x))$ for the predicted dominant/complement allocation, and $\boldsymbol{\beta}_\phi(x) = (\beta_{\phi,0}(x), \ldots, \beta_{\phi,L-1}(x))$ for the predicted depth allocation across the L nonlinear blocks. The scalar $T_\beta > 0$ is the fixed depth-softmax temperature. The symbol σ denotes the logistic sigmoid, and $\mathrm{softmax}$ denotes the standard normalization that maps a vector of logits to a probability simplex; w_ρ and b_ρ are the intensity-head weight vector and bias; W_π and b_π are the subspace-allocation head parameters; and W_β and b_β are the depth-allocation head parameters. GANR computes

$$g_\phi(x) = C_\phi(x, e_R, e_\perp), \qquad \rho_\phi(x) = \sigma(w_\rho^\top g_\phi + b_\rho),$$
$$\boldsymbol{\pi}_\phi(x) = \mathrm{softmax}(W_\pi g_\phi + b_\pi), \qquad \boldsymbol{\beta}_\phi(x) = \mathrm{softmax}((W_\beta g_\phi + b_\beta)/T_\beta)\,. \tag{10}$$

The first two outputs are supervised toward the NRGD intensity and subspace-allocation targets. The depth-allocation vector has no physical label because neural depth is a representation choice

rather than a property of the physical system; imposing a synthetic physical target on depth would mix system geometry with architecture design. It is therefore learned only through the end-to-end predictive objective.

Layer-wise geometry modulation. For nonlinear-block index $\ell \in \{0, \dots, L-1\}$, each subspace code is mapped into the layer preactivation coordinates before the two branches are combined. For branch $b \in \{R, \perp\}$, let $H_{\ell,b}$ denote the fixed highway projection and $U_{\ell,b}V_{\ell,b}^\top$ a trainable low-rank residual projection. We define $\mathcal{P}_{\ell,b}(e_b) = (H_{\ell,b} + U_{\ell,b}V_{\ell,b}^\top)e_b$. Let $u_\ell(x)$ denote the baseline nonlinear preactivation, let $\mathrm{RMS}(\cdot)$ denote root-mean-square magnitude, and let $\epsilon_A > 0$ be a numerical floor. The raw geometry direction $a_\ell^{\mathrm{geo}}(x)$ and the scale-matched shift $\widetilde{a}_\phi^{(\ell)}(x)$ are

$$\begin{aligned} a_\ell^{\mathrm{geo}}(x) &= \sqrt{\pi_{\phi,R}(x)}\,\mathcal{P}_{\ell,R}(e_R(x)) + \sqrt{\pi_{\phi,\perp}(x)}\,\mathcal{P}_{\ell,\perp}(e_\perp(x)), \\ \widetilde{a}_\phi^{(\ell)}(x) &= \rho_\phi(x)\sqrt{\beta_{\phi,\ell}(x)}\,\frac{a_\ell^{\mathrm{geo}}(x)}{\mathrm{RMS}(a_\ell^{\mathrm{geo}}(x)) + \epsilon_A}\,\mathrm{RMS}(u_\ell(x)). \end{aligned} \tag{11}$$

The square-root factors provide an amplitude parameterization for the probability-like controls, but the two projected branches need not have equal norms or remain orthogonal after projection. We therefore interpret $\boldsymbol{\pi}_\phi$ as physically informed dominant/complement branch-mixing weights rather than as an exact post-projection energy split. RMS matching controls the overall layer scale, ρ_ϕ controls local modulation strength, and $\boldsymbol{\beta}_\phi$ routes the modulation across depth while leaving the activation itself unchanged. Appendix A.3 gives the exact dimensions and low-rank factorization.

Backbone interface. For a residual backbone (He et al., 2016), let $h^{(\ell)}$ denote the hidden state entering block ℓ. Root-mean-square normalization, written as $\mathrm{RMSNorm}$, is applied before each nonlinear preactivation (Zhang & Sennrich, 2019). For a single-branch activation, W_ℓ and b_ℓ denote the preactivation weight matrix and bias, u_ℓ denotes the baseline preactivation, and $\widetilde{u}_\ell$ denotes its GA-NAR-modulated counterpart. The scalar α_ℓ is the adapter scale for layer ℓ. The output projection matrix and bias are denoted by $W_{o,\ell}$ and $b_{o,\ell}$, and φ_B denotes the activation function associated with backbone family B. The block is

$$\begin{aligned} u_\ell = W_\ell\,\mathrm{RMSNorm}(h^{(\ell)}) + b_\ell, \qquad \widetilde{u}_\ell = u_\ell + \alpha_\ell \widetilde{a}_\phi^{(\ell)}, \\ h^{(\ell+1)} = h^{(\ell)} + W_{o,\ell}\varphi_B(\widetilde{u}_\ell) + b_{o,\ell}. \end{aligned} \tag{12}$$

The chosen activation family φ_B is one of ReLU, GELU, SiLU, Mish, or Softplus. GELU and SiLU provide smooth modern alternatives to ReLU (Hendrycks & Gimpel, 2016; Ramachandran et al., 2018), while Mish supplies a non-monotonic smooth activation (Misra, 2020). A broader activation-function review motivates treating the activation as a controlled backbone factor rather than as part of GA-NAR itself (Apicella et al., 2021).

For the SwiGLU backbone (Shazeer, 2020), $W_{g,\ell}$ and $b_{g,\ell}$ denote the gate-branch weight matrix and bias, while $W_{v,\ell}$ and $b_{v,\ell}$ denote the value-branch parameters. Their baseline preactivations are denoted by g_ℓ and v_ℓ, and $\widetilde{g}_\ell$ denotes the modulated gate preactivation. The symbol $\odot$ denotes element-wise multiplication. GANR modulates only the gate branch,

$$\begin{aligned} g_\ell = W_{g,\ell}\,\mathrm{RMSNorm}(h^{(\ell)}) + b_{g,\ell}, \qquad v_\ell = W_{v,\ell}\,\mathrm{RMSNorm}(h^{(\ell)}) + b_{v,\ell}, \\ \widetilde{g}_\ell = g_\ell + \alpha_\ell \widetilde{a}_\phi^{(\ell)}, \qquad h^{(\ell+1)} = h^{(\ell)} + W_{o,\ell}[\mathrm{SiLU}(\widetilde{g}_\ell) \odot v_\ell] + b_{o,\ell}. \end{aligned} \tag{13}$$

The layer-wise mechanism therefore changes the operating point of the nonlinear transformation without redefining the activation itself. Setting every modulation to zero recovers the original backbone exactly.

Training objective. The total GA-NAR training loss is denoted by $\mathcal{L}$. Besides the prediction term $\mathcal{L}_{\mathrm{pred}}$ introduced in Section 3, we use $\mathcal{L}_{\mathrm{int}}$ for intensity-target supervision and $\mathcal{L}_{\mathrm{alloc}}$ for subspace-allocation supervision. Their nonnegative weights are denoted by λ_ρ and λ_π. The function $H^2(p, q)$ denotes squared Hellinger distance between probability vectors p and q, and the expectations in the two auxiliary losses are taken over their fixed training-only geometry-target sets. The objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda_\rho \mathcal{L}_{\mathrm{int}} + \lambda_\pi \mathcal{L}_{\mathrm{alloc}}, \qquad \mathcal{L}_{\mathrm{int}} = \mathbb{E}[(\rho_\phi - \rho_G^\star)^2], \qquad \mathcal{L}_{\mathrm{alloc}} = \mathbb{E}[H^2(\boldsymbol{\pi}_\phi, \boldsymbol{\pi}_G^\star)]. \tag{14}$$

The loss weights are selected only in the development protocol and are then frozen. No depth target and no depth-specific loss are introduced.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Experiments follow the method hierarchy: NRGD is tested independently as a descriptor interface, GANR is tested under matched auxiliary-physics conditions, and GA-NAR is evaluated end to end. Benchmark definitions and complete protocols are given in Appendices E and F.

Table 1: NRGD descriptor prediction across five systems (mean ± SD, three seeds; lower is better).

System	Method	Allocation MSE ↓ ($\times 10^{-3}$)	Total-response MSE ↓ ($\times 10^{-3}$)	Fraction NMSE ↓ dimensionless
Synthetic	NRGD (input-only)	1.901±0.014	7.707±0.090	**0.890±0.037**
	NRGD + directional sketch	**0.241±0.002**	**0.736±0.004**	1.000±0.015
	Raw-input control	4.848±0.056	19.735±0.393	2.507±0.414
	Random-projector control	4.781±0.141	18.645±0.196	2.720±0.253
	Intensity-only control	22.995±0.356	18.515±1.245	319.281±15.640
	Training-mean control	4.723±0.000	19.054±0.000	1.006±0.000
IEEE-118	NRGD (input-only)	0.669±0.011	3.355±0.042	1.068±0.040
	NRGD + directional sketch	**0.168±0.000**	**0.487±0.003**	**0.952±0.004**
	Raw-input control	0.908±0.020	4.044±0.168	2.218±0.157
	Random-projector control	0.880±0.008	4.358±0.032	1.254±0.056
	Intensity-only control	14.914±0.037	3.800±0.177	155.312±14.930
	Training-mean control	2.157±0.000	10.577±0.000	1.008±0.000
Duffing	NRGD (input-only)	0.248±0.003	1.544±0.034	1.008±0.002
	NRGD + directional sketch	**0.132±0.000**	**0.281±0.005**	**0.855±0.000**
	Raw-input control	0.267±0.004	1.597±0.022	1.116±0.014
	Random-projector control	0.264±0.011	1.624±0.082	1.087±0.036
	Intensity-only control	0.720±0.008	1.839±0.230	4.892±0.249
	Training-mean control	0.266±0.000	1.776±0.000	1.004±0.000
Allen–Cahn	NRGD (input-only)	0.079±0.001	0.776±0.011	1.074±0.039
	NRGD + directional sketch	**0.033±0.000**	**0.150±0.003**	**0.993±0.003**
	Raw-input control	0.109±0.002	1.056±0.029	1.800±0.235
	Random-projector control	0.115±0.009	1.109±0.081	1.240±0.047
	Intensity-only control	7.740±0.168	1.135±0.051	398.857±7.798
	Training-mean control	1.498±0.000	15.007±0.000	1.002±0.000
Shallow-water	NRGD (input-only)	0.030±0.000	0.278±0.005	1.080±0.020
	NRGD + directional sketch	**0.021±0.000**	**0.110±0.001**	**0.927±0.001**
	Raw-input control	0.042±0.002	0.328±0.015	1.754±0.024
	Random-projector control	0.038±0.002	0.331±0.032	1.309±0.026
	Intensity-only control	6.774±0.028	0.332±0.022	503.327±22.674
	Training-mean control	0.447±0.000	5.233±0.000	1.002±0.000

5.2 NRGD: INDEPENDENT RESPONSE-GEOMETRY INTERFACE

Table 1 tests whether the NRGD descriptors can be recovered independently of the final predictor; exact five-system benchmark specifications are given in Appendix E.1, and final allocation-only hyperparameters are summarized in Table 8. On every system, NRGD (input-only) has lower allocation and total-response MSE than each input-only control, and the directional-response sketch reduces both errors further. Thus the aggregate intensity/allocation information used by GANR is recoverable from input-side information; the less consistent Fraction NMSE identifies fine directional fractions as the harder target. Appendix B then tests whether the underlying response geometry remains stable across finite scales, spectra, probe budgets, and operating states.

5.3 GANR: CONTROLLED RESPONSE-AWARE REPRESENTATION

To separate response-aware nonlinear representation from privileged auxiliary physics, Table 2 evaluates the same model set under No physics and Shared physics conditions; the Controlled implicit system is defined in Appendix E.2. GANR attains the lowest NRMSE on Controlled, Duffing, IEEE-118, and shallow-water in both conditions; Allen–Cahn is the consistent exception. The same four-of-five pattern when the equation-derived linearization is either absent from every model or shared by every model shows that the advantage is not explained by unequal auxiliary-physics access. Appendix C tests the same premise with response stratification and a stricter shared plug-in interface.

Table 2: GANR factorial NRMSE ($\times 10^{-3}$, mean $\pm$ SD over three seeds; lower is better). For each model, the upper row is No physics and the lower row is Shared physics. Best and second-best values within each condition are bold-underlined and underlined.

Model	Controlled	Duffing	IEEE-118	Allen–Cahn	Shallow-water
GANR	**5.20±0.03** **5.60±0.12**	**3.27±0.05** **2.70±0.07**	**6.59±0.03** **0.54±0.05**	12.48±0.02 11.85±0.01	**6.87±0.02** **5.83±0.02**
ReLU	12.43±0.07 11.74±0.05	9.49±0.05 7.69±0.08	10.25±0.05 3.40±0.02	15.41±0.03 13.10±0.01	13.24±0.09 8.44±0.01
GELU	9.68±0.11 9.02±0.04	7.99±0.04 5.52±0.00	10.18±0.09 2.97±0.01	14.37±0.04 12.07±0.00	11.01±0.10 6.43±0.02
SiLU	12.24±0.13 11.62±0.19	10.10±0.01 6.01±0.01	10.06±0.06 2.61±0.01	14.63±0.05 12.55±0.01	11.00±0.12 5.96±0.02
SwiGLU	7.81±0.36 8.60±0.16	4.89±0.12 2.72±0.02	7.82±0.06 1.16±0.01	12.27±0.03 **10.56±0.03**	12.09±0.08 11.80±0.01
ResMLP	7.57±0.55 7.62±0.40	4.87±0.30 2.87±0.12	7.41±0.09 1.34±0.01	**12.17±0.05** 12.14±0.17	11.26±0.03 10.13±0.08
Fourier	23.90±0.67 20.63±0.58	20.23±0.35 16.80±0.26	14.24±0.49 2.22±0.07	16.57±0.09 14.70±0.08	18.16±0.04 18.08±0.02

5.4 GA-NAR: END-TO-END EVALUATION

Table 3: Paired end-to-end results (mean $\pm$ SD over three seeds). MSE reduction is relative to the activation-matched baseline; all 18 reductions are positive. Lower is better for MSE and residual.

System	Activation	Baseline MSE ($\times 10^{-4}$)	GA-NAR MSE ($\times 10^{-4}$)	MSE reduction ↑ (%)	Baseline residual ($\times 10^{-3}$)	GA-NAR residual ($\times 10^{-3}$)
IEEE-118 AC power flow	ReLU	0.0259±0.0021	0.0238±0.0018	8.1	8.0308±0.5511	6.9492±0.3613
	GELU	0.0399±0.0007	0.0353±0.0007	11.5	10.6149±0.1277	9.2665±0.1581
	SiLU	0.0435±0.0012	0.0381±0.0014	12.4	11.3935±0.2659	9.7841±0.4621
	Mish	0.0381±0.0013	0.0337±0.0012	11.5	10.5189±0.2380	8.9520±0.3089
	Softplus	0.0471±0.0023	0.0449±0.0024	4.7	9.8486±0.0706	9.5740±0.1361
	SwiGLU	0.0182±0.0008	0.0144±0.0006	20.9	8.5458±0.0817	6.2811±0.2028
Kuramoto-64	ReLU	186.7747±2.5028	10.6688±1.4531	94.3	39.1765±2.2247	27.9715±2.8399
	GELU	3.2497±0.4064	3.1903±0.4001	1.8	12.0207±0.0825	11.9496±0.0643
	SiLU	2.2618±0.0834	2.2146±0.0736	2.1	11.6913±0.1073	11.6352±0.0895
	Mish	2.4429±0.0922	2.3912±0.0883	2.1	11.9841±0.0239	11.8982±0.0326
	Softplus	3.8716±0.0707	3.8198±0.0695	1.3	12.4950±0.0709	12.4272±0.0590
	SwiGLU	236.0363±7.0771	47.0842±0.8463	80.1	68.2155±5.7656	64.4192±5.2922
Nonlinear heat FEM	ReLU	0.0239±0.0007	0.0234±0.0007	2.1	0.0401±0.0011	0.0402±0.0011
	GELU	0.0448±0.0008	0.0436±0.0009	2.7	0.0647±0.0010	0.0643±0.0010
	SiLU	0.0418±0.0014	0.0406±0.0014	2.9	0.0624±0.0010	0.0620±0.0010
	Mish	0.0465±0.0019	0.0450±0.0018	3.2	0.0653±0.0009	0.0648±0.0009
	Softplus	0.0446±0.0006	0.0148±0.0005	66.8	0.0720±0.0005	0.1169±0.0033
	SwiGLU	0.0173±0.0010	0.0170±0.0010	1.7	0.0378±0.0009	0.0379±0.0009

Table 3 tests GA-NAR across six activation families and three physical systems. Standardized test MSE decreases in all 18 activation–system pairs and in every paired seed. Median reductions are 11.5%, 2.1%, and 2.8% on IEEE-118, Kuramoto-64, and heat FEM; the larger means on the latter two systems come from a few large activation-specific gains. Physical residuals decrease for every activation on IEEE-118 and Kuramoto-64, whereas heat FEM exposes prediction–residual trade-offs, most clearly for Softplus. Because GA-NAR adds a correction branch, Appendix H matches correction capacity and NRGD access in a four-way ablation. Full GA-NAR gives larger paired MSE/residual improvements than Capacity Control in 5/7 settings and Generic Geometry Control in 6/7; neither confound alone explains the overall gain, although the controls win in isolated cases.

6 CONCLUSION

GA-NAR aligns neural nonlinear computation with heterogeneous physical response. NRGD identifies a recurring dominant-response subspace plus state-dependent intensity/allocation, and GANR realizes them through dual-subspace, scale-matched, layer-wise modulation. All 18 MSE comparisons improve. The controlled ablation (Appendix H) shows that capacity and generic geometry access contribute but do not fully explain the pattern, while structured modulation is not universally best. Limitations include finite-scale/coordinate dependence, three seeds, and benchmark scope.

REFERENCES

Andrea Apicella, Francesco Donnarumma, Francesco Isgrò, and Roberto Prevete. A survey on modern trainable activation functions. *Neural Networks*, 138:14–32, 2021. doi: 10.1016/j.neunet.2021.01.026.

Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966. doi: 10.2140/pjm.1966.16.1.

Abdolmehdi Behroozi, Chaopeng Shen, and Daniel Kifer. Sensitivity-constrained Fourier Neural Operators for forward and inverse problems in parametric differential equations. In *International Conference on Learning Representations*, 2025.

Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, 3 edition, 2008. doi: 10.1007/978-0-387-75934-0.

Qianying Cao, Somdatta Goswami, and George Em Karniadakis. Laplace Neural Operator for solving differential equations. *Nature Machine Intelligence*, 6:631–640, 2024. doi: 10.1038/s42256-024-00844-4.

Peng Chen and Omar Ghattas. Hessian-based sampling for high-dimensional model reduction. *International Journal for Uncertainty Quantification*, 9(2):103–121, 2019. doi: 10.1615/Int.J.UncertaintyQuantification.2019028753.

Wuyang Chen, Jialin Song, Pu Ren, Shashank Subramanian, Dmitriy Morozov, and Michael W. Mahoney. Data-efficient operator learning via unsupervised pretraining and in-context learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 6213–6245, 2024. doi: 10.52202/079017-0201.

Ze Cheng, Zhongkai Hao, Xiaoqiang Wang, Jianing Huang, Youjia Wu, Xudan Liu, Yiru Zhao, Songming Liu, and Hang Su. Reference neural operators: Learning the smooth dependence of solutions of PDEs on geometric deformations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8060–8076, 2024.

Ze Cheng, Zhuoyu Li, Xiaoqiang Wang, Jianing Huang, Zhizhou Zhang, Zhongkai Hao, and Hang Su. Accelerating PDE-constrained optimization by the derivative of neural operators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10090–10105, 2025.

Namkyeong Cho, Junseung Ryu, and Hyung Ju Hwang. Sobolev training for operator learning. *Journal of Computational Physics*, 543:114408, 2025. doi: 10.1016/j.jcp.2025.114408.

Woojin Cho, Minju Jo, Haksoo Lim, Kookjin Lee, Dongeun Lee, Sanghyun Hong, and Noseong Park. Parameterized physics-informed neural networks for parameterized PDEs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8510–8533, 2024.

Paul G. Constantine. *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. SIAM, 2015. doi: 10.1137/1.9781611973860.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4278–4287, 2017.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279, 2008. doi: 10.1145/1390156.1390191.

Francisco Eiras, Adel Bibi, Rudy R. Bunel, Krishnamurthy Dj Dvijotham, Philip Torr, and M. Pawan Kumar. Efficient error certification for physics-informed neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12318–12347, 2024.

J. Gonzalez-Sieiro, D. Pardo, and V. M. Calo. Efficient operator learning with derivative-enhanced parameter sensitivity information and hybrid optimization. *Engineering with Computers*, 42:109, 2026. doi: 10.1007/s00366-026-02327-z.

Vignesh Gopakumar, Ander Gray, Daniel Giles, Lorenzo Zanisi, Matt J. Kusner, Timo Betcke, Stanislas Pamela, and Marc Peter Deisenroth. Learning physical operators using neural operators. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3223–3231, 2026.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. DPOT: Auto-regressive denoising operator transformer for large-scale PDE pre-training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17616–17635, 2024a.

Zhongkai Hao, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, and Jun Zhu. PINNacle: A comprehensive benchmark of physics-informed neural networks for solving PDEs. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-2442.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 19(2):433–450, 1990. doi: 10.1080/03610919008812866.

Ameya D. Jagtap, Kenji Kawaguchi, and George Em Karniadakis. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404:109136, 2020a. doi: 10.1016/j.jcp.2019.109136.

Ameya D. Jagtap, Kenji Kawaguchi, and George Em Karniadakis. Locally adaptive activation functions with slope recovery for deep and physics-informed neural networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2239):20200334, 2020b. doi: 10.1098/rspa.2020.0334.

Katiana Kontolati, Somdatta Goswami, George Em Karniadakis, and Michael D. Shields. Learning nonlinear operators in latent spaces for real-time predictions of complex dynamics in physical systems. *Nature Communications*, 15:5101, 2024. doi: 10.1038/s41467-024-49411-w.

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.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 26548–26560, 2021.

Yoshiki Kuramoto. *Chemical Oscillations, Waves, and Turbulence*, volume 19 of *Springer Series in Synergetics*. Springer, 1984. doi: 10.1007/978-3-642-69689-3.

Jongmok Lee, Chaeyun Won, Anna Lee, Bumsoo Park, Sooyoung Lee, and Seungchul Lee. A geometry-adaptive physics-informed operator framework generalized for arbitrary geometries. *Engineering Applications of Artificial Intelligence*, 174:114457, 2026. doi: 10.1016/j.engappai.2026.114457.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research*, 2023a.

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, 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, 2023b.

Levi E. Lingsch, Mike Yan Michelis, Emmanuel De Bezenac, Sirani M. Perera, Robert K. Katzschmann, and Siddhartha Mishra. Beyond regular grids: Fourier-based neural operators on arbitrary domains. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30610–30629, 2024.

Ning Liu, Yiming Fan, Xianyi Zeng, Milan Klöwer, Lu Zhang, and Yue Yu. Harnessing the power of neural operators with automatically encoded conservation laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30965–30997, 2024.

Shouyi Liu, Xiaokang Yang, and Yuntian Chen. Riesz Neural Operator for solving partial differential equations. In *International Conference on Learning Representations*, 2026.

Miguel Liu-Schiaffini, Julius Berner, Boris Bonev, Thorsten Kurth, Kamyar Azizzadenesheli, and Anima Anandkumar. Neural operators with localized integral and differential kernels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32576–32594, 2024.

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*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Dingcheng Luo, Thomas O'Leary-Roseberry, Peng Chen, and Omar Ghattas. Dimension reduction for derivative-informed operator learning: An analysis of approximation errors. *Journal of Machine Learning Research*, 27(182):1–94, 2026.

Emilia Magnani, Marvin Pförtner, Tobias Weber, and Philipp Hennig. Linearization turns neural operators into function-valued Gaussian processes. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42421–42452, 2025.

Diganta Misra. Mish: A self regularized non-monotonic neural activation function. In *British Machine Vision Conference*, 2020.

Thomas O'Leary-Roseberry, Umberto Villa, Peng Chen, and Omar Ghattas. Derivative-informed projected neural networks for high-dimensional parametric maps governed by PDEs. *Computer Methods in Applied Mechanics and Engineering*, 388:114199, 2022. doi: 10.1016/j.cma.2021.114199.

Thomas O'Leary-Roseberry, Peng Chen, Umberto Villa, and Omar Ghattas. Derivative-Informed Neural Operator: An efficient framework for high-dimensional parametric derivative learning. *Journal of Computational Physics*, 496:112555, 2024. doi: 10.1016/j.jcp.2023.112555.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Yuan Qiu, Nolan Bridges, and Peng Chen. Derivative-enhanced Deep Operator Network. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0660.

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. In *6th International Conference on Learning Representations (ICLR 2018), Workshop Track Proceedings*, 2018.

Alexander Rudikov, Vladimir Fanaskov, Ekaterina Muravleva, Yuri M. Laevsky, and Ivan Oseledets. Neural operators meet conjugate gradients: The FCG-NO method for efficient PDE solving. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42766–42782, 2024.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Zakhar Shumaylov, Peter Zaika, James Rowbottom, Ferdia Sherry, Melanie Weber, and Carola-Bibiane Schönlieb. Lie Algebra Canonicalization: Equivariant neural operators under arbitrary Lie Groups. In *International Conference on Learning Representations*, 2025.

Ashutosh Singh, Ricardo Augusto Borsoi, Deniz Erdogmus, and Tales Imbiriba. Learning semi-linear neural operators: A unified recursive framework for prediction and data assimilation. In *International Conference on Learning Representations*, 2024.

Steven H. Strogatz. From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1–4):1–20, 2000. doi: 10.1016/S0167-2789(00)00094-4.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1741.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022. doi: 10.1016/j.jcp.2021.110768.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998. doi: 10.1038/30918.

Zipeng Xiao, Zhongkai Hao, Bokai Lin, Zhijie Deng, and Hang Su. Improved operator learning by orthogonal attention. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54288–54299, 2024.

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

Ray D. Zimmerman, Carlos E. Murillo-Sanchez, and Robert J. Thomas. MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, 2011. doi: 10.1109/TPWRS.2010.2051168.

AI USE STATEMENT

Generative AI tools assisted with literature search, mathematical and implementation checks, experimental-design feedback, and manuscript drafting and editing. Every AI-assisted statement, citation, derivation, protocol choice, and implementation detail was independently reviewed by the authors against the governing definitions, primary sources, and numerical results. Generative AI was not used as empirical evidence and was not used to fabricate, alter, or selectively report experimental observations. The authors take full responsibility for the final manuscript and supplementary materials.

Appendix Contents

Supplementary Material for GA-NAR

A Notation, Preprocessing, and Implementation Details – – – – – – – – – – – – – – – – **16**

A.1 Standardization, residual scaling, and implicit JVPs – – – – – – – – – – – – – – – – **16**

A.2 Spectral projection and rank selection in NRGD – – – – – – – – – – – – – – – – **16**

A.3 Backbone-compatible GANR adapter and target losses – – – – – – – – – – – – – – – – **17**

B Additional NRGD Experimental Analysis – **19**

B.1 Diagnostic Quantities and Experimental Settings – – – – – – – – – – – – – – – – **19**

B.2 Finite-Scale Stability and Global Response Structure – – – – – – – – – – – – – – – – **20**

B.3 Sensitivity to Finite Geometry Budgets – **21**

B.4 State-Dependent Local Response Structure – **22**

B.5 Extended Interface-Fidelity Analysis – **23**

B.6 Summary and Scope – **24**

C Independent Evidence for GANR Geometry Conditioning – – – – – – – – – – – – – – – – **24**

C.1 Geometry Interface and Evidence Boundary – **25**

C.2 Controlled GANR Experimental Settings – **26**

C.3 Response-Conditioned Predictive Difficulty – **26**

C.4 Response-Aware Nonlinear Representation Under Controlled Settings – – – – – – – – **27**

C.5 Strict Plug-In Test of Response-Aware Nonlinearity – – – – – – – – – – – – – – – – **27**

C.6 What the Independent Evidence Supports About GANR – – – – – – – – – – – – – – – – **28**

D Theoretical Analysis and Proofs – **28**

D.1 Proposition 1: finite-sample non-identifiability of nonlinear-response geometry – – – – – **28**

D.2 Proposition 2: smooth-limit Hessian–Gram response identity – – – – – – – – – – – – **29**

D.3 Spherical moment identity – **29**

D.4 Trace decomposition and basis invariance of the dominant/complement allocation – – – – **29**

D.5 Proposition 3: scale control and exact baseline recovery of the layer-wise adapter – – – – **29**

D.6 Analytic quartic oracle – **30**

E Benchmark Definitions – **30**

E.1 Formal five-system NRGD benchmark specifications – – – – – – – – – – – – – – – – **30**

E.2 Analytical and auxiliary synthetic benchmarks – – – – – – – – – – – – – – – – – – **31**

F Experiment-specific Protocols – **33**

F.1 NRGD interface-fidelity protocol – **33**

F.2 GANR factorial response-representation protocol – – – – – – – – – – – – – – – – – – **33**

F.3 GA-NAR end-to-end protocol – **34**

G Backbone Pairing and Capacity Context – **34**

H Controlled Attribution of GA-NAR Gains – **34**

I Reported Metrics and State-Dependence Diagnostics – – – – – – – – – – – – – – – – – – **36**

J Prespecified Experimental Protocol – **37**

A Notation, Preprocessing, and Implementation Details

Symbols defined in the main paper retain exactly the same meaning throughout the appendix. In particular, the standardized input and output are denoted by x and y, the physical parameter and state by q and z, the local parameter-to-solution map by G, and the NRGD projector by P_R. The appendix introduces additional implementation symbols only when they are needed, and each new symbol is defined in prose before it enters a displayed equation.

A.1 Standardization, residual scaling, and implicit JVPs

All physical parameters and states are standardized using statistics computed only from the supervised training split. Let $\mu_q \in \mathbb{R}^d$ and $\mu_z \in \mathbb{R}^m$ denote the training means of the physical parameter and state, respectively. Let $s_{q,i} > 0$ and $s_{z,j} > 0$ denote the retained coordinate-wise sample standard deviations, and let $D_q \in \mathbb{R}^{d\times d}$ and $D_z \in \mathbb{R}^{m\times m}$ be the corresponding diagonal scaling matrices. Constant coordinates are removed before training, so every retained scale is strictly positive. This convention fixes the metric in which prediction error and nonlinear-response geometry are measured and follows standard normalization practice in neural-operator implementations (Kovachki et al., 2023). The complete coordinate transformation is

$$\mu_q = (\mu_{q,1}, \dots, \mu_{q,d}), \qquad \mu_z = (\mu_{z,1}, \dots, \mu_{z,m}), \tag{15}$$

$$D_q = \operatorname{diag}(s_{q,1}, \dots, s_{q,d}), \qquad D_z = \operatorname{diag}(s_{z,1}, \dots, s_{z,m}), \tag{16}$$

$$q = \mu_q + D_q x, \qquad z = \mu_z + D_z y. \tag{17}$$

The same transformation is used for supervised data, physical residuals, derivative calculations, and geometry probes. As a result, every response magnitude reported by NRGD is relative to this fixed train-derived coordinate system; it is not interpreted as an intrinsic coordinate-free curvature.

Residual scaling is introduced separately from input/output standardization. Let F° denote the governing residual after the physical variables have been expressed through standardized coordinates. Let $(y_{\mathrm{ref}}, x_{\mathrm{ref}})$ be the nominal regular reference state, and let A°_{ref} be the state Jacobian of F° at that reference point. For residual row i, let e_i be the corresponding Euclidean basis vector, let $s_{F,i}$ be the row scale, and let D_F be the diagonal matrix of all row scales. These objects define the scaled residual $\bar{F}$ used throughout numerical evaluation:

$$F^\circ(y, x) = \mathcal{F}(\mu_z + D_z y, \mu_q + D_q x), \qquad A^\circ_{\mathrm{ref}} = \frac{\partial F^\circ}{\partial y}(y_{\mathrm{ref}}, x_{\mathrm{ref}}), \tag{18}$$

$$s_{F,i} = \max\left\{ \|e_i^\top A^\circ_{\mathrm{ref}}\|_2, 10^{-12} \right\}, \qquad D_F = \operatorname{diag}(s_{F,1}, \dots, s_{F,m}), \tag{19}$$

$$\bar{F}(y, x) = D_F^{-1} F^\circ(y, x). \tag{20}$$

Left scaling changes numerical conditioning and the residual norm used for diagnostics, but it does not change the exact local solution derivative. Derivative-informed surrogate methods similarly rely on stable derivative actions rather than dense Jacobian materialization (O'Leary-Roseberry et al., 2022).

For matrix-free differentiation, let $v \in \mathbb{R}^d$ be an arbitrary standardized input direction and let $u \in \mathbb{R}^m$ denote the corresponding directional derivative of the standardized solution map. The matrices $\bar{F}_x$ and $\bar{F}_y$ denote the parameter and state Jacobians of the scaled residual. The implicit Jacobian–vector product is obtained from one linear solve,

$$\bar{F}_y(y, x)u = -\bar{F}_x(y, x)v, \qquad u = J_G(x)v. \tag{21}$$

This is the derivative evaluation used by NRGD. It matches the Jacobian-action viewpoint used in derivative-informed operator learning (O'Leary-Roseberry et al., 2024); dense solution Jacobians and Hessians are never assembled during geometry construction.

A.2 Spectral projection and rank selection in NRGD

NRGD converts the normalized directional-response moment into a positive-semidefinite trace-one operator before extracting the dominant response subspace. Let $\widehat{\Gamma}_G \in \mathbb{R}^{d\times d}$ denote the normalized

empirical response moment from the main method. Let I_d be the $d \times d$ identity. The finite-scale backprojected operator is denoted by $\widehat{K}_G$. Its eigendecomposition uses the orthogonal matrix V and raw eigenvalue vector $\kappa \in \mathbb{R}^d$. Because finite-scale measurements need not yield a nonnegative spectrum, the raw eigenvalues are projected onto the probability simplex, a standard Euclidean projection problem (Duchi et al., 2008). The projected vector is denoted by $\widetilde{\lambda}$, the symbol $\mathbf{1}$ denotes the all-ones vector, and σ denotes the permutation that sorts the projected entries in nonincreasing order. The resulting ordered eigenvalues and eigenvectors are $\lambda^\star$ and $V^\star$. The projected response operator $\widehat{M}_G$ and selected response rank R are then defined by

$$\widehat{K}_G = \frac{(d+2)\widehat{\Gamma}_G - I_d}{2}, \qquad \widehat{K}_G = V \operatorname{diag}(\kappa) V^\top, \tag{22}$$

$$\widetilde{\lambda} = \arg\min_{\lambda \in \mathbb{R}^d} \|\lambda - \kappa\|_2^2, \qquad \lambda_k \geq 0, \qquad \mathbf{1}^\top \lambda = 1, \tag{23}$$

$$\lambda_k^\star = \widetilde{\lambda}_{\sigma(k)}, \qquad V^\star = [v_{\sigma(1)}, \ldots, v_{\sigma(d)}], \tag{24}$$

$$\widehat{M}_G = V^\star \operatorname{diag}(\lambda^\star) V^{\star\top}, \tag{25}$$

$$R = \min\left\{ r : \sum_{k=1}^{r} \lambda_k^\star \geq 0.90 \right\}. \tag{26}$$

The simplex projection is used only to obtain a stable finite-scale positive-semidefinite operator for response-rank selection; it is not claimed to recover an exact Hessian. The cumulative 90% rule therefore defines a reproducible empirical response rank rather than a population-optimal intrinsic dimension.

The object passed to GANR is the dominant subspace projector, not a particular eigenbasis. The affine backprojection and eigenvalue-only simplex projection preserve the eigenvectors of $\widehat{\Gamma}_G$. Let U be its orthogonal eigenvector matrix, let γ denote the corresponding eigenvalues ordered from largest to smallest, and let $U_{1:R} \in \mathbb{R}^{d \times R}$ collect the leading R eigenvectors. The projector is P_R. The constants K_H, K_S, K_R, and $K_\perp$ denote the fixed numbers of Hutchinson probes, global sphere directions, dominant-subspace directions, and complementary-subspace directions, respectively. Their definitions are

$$\widehat{\Gamma}_G = U \operatorname{diag}(\gamma) U^\top, \qquad \gamma_1 \geq \ldots \geq \gamma_d, \tag{27}$$

$$U_{1:R} \in \mathbb{R}^{d \times R}, \qquad P_R = U_{1:R} U_{1:R}^\top, \qquad K_H = K_S = 8, \quad K_R = K_\perp = 4. \tag{28}$$

The Frobenius-response estimator uses the Hutchinson construction (Hutchinson, 1990). Because downstream computation depends on P_R rather than on individual eigenvectors, any orthogonal rotation inside the retained eigenspace leaves the decomposition unchanged.

A.3 Backbone-compatible GANR adapter and target losses

GANR uses one neural interface across all evaluated activation families. The frozen projector produces the dominant and complementary standardized inputs x_R and $x_\perp$. Their encoders output branch representations $e_R(x)$ and $e_\perp(x)$ in an adapter space of width d_A. The number of residual blocks is L, and the shared controller width is d_{ctrl}. The controller feature is denoted by $g_{\mathrm{ctrl}}(x)$. Its three heads produce the scalar intensity ρ_ϕ, the two-component subspace allocation $\boldsymbol{\pi}_\phi$, and the L-component depth allocation $\boldsymbol{\beta}_\phi$. The matrices and biases with subscripts ρ, π, and β are the trainable head parameters, while $T_\beta > 0$ is the fixed depth-softmax temperature. Using Δ^k for the probability simplex with $k+1$ components, the controller is specified by

$$x_R = P_R x, \qquad x_\perp = (I_d - P_R)x, \qquad d_A = 64, \qquad L = 4, \qquad d_{\mathrm{ctrl}} = 32, \tag{29}$$

$$\rho_\phi(x) = \operatorname{sigmoid}(w_\rho^\top g_{\mathrm{ctrl}}(x) + b_\rho), \tag{30}$$

$$\boldsymbol{\pi}_\phi(x) = \operatorname{softmax}(W_\pi g_{\mathrm{ctrl}}(x) + b_\pi), \tag{31}$$

$$\boldsymbol{\beta}_\phi(x) = \operatorname{softmax}((W_\beta g_{\mathrm{ctrl}}(x) + b_\beta)/T_\beta), \tag{32}$$

$$\boldsymbol{\pi}_\phi(x) \in \Delta^1, \qquad \boldsymbol{\beta}_\phi(x) \in \Delta^{L-1}. \tag{33}$$

The intensity and subspace-allocation heads are supervised by NRGD targets. The depth head is optimized only through prediction because the physical geometry does not define a privileged neural depth. RMS normalization is shared by the baseline and augmented backbones (Zhang & Sennrich, 2019), while the residual pathway follows the standard residual-network construction (He et al., 2016). For SwiGLU, the gate/value decomposition follows the gated feed-forward design (Shazeer, 2020).

Each layer maps the shared branch codes into its own nonlinear-preactivation coordinates. For layer ℓ and branch $b \in \{R, \perp\}$, let p_ℓ be the target preactivation width, $H_{\ell,b} \in \mathbb{R}^{p_\ell \times d_A}$ the fixed highway, and $U_{\ell,b}V_{\ell,b}^\top$ the trainable rank-r_A residual projection. The vector $r_{\ell,b}(x)$ is the resulting branch representation in layer coordinates. The raw geometry direction $a_\ell^{\mathrm{geo}}(x)$ combines the two branches using the predicted subspace allocation. The function $\mathrm{RMS}(\cdot)$ denotes root-mean-square magnitude, with w used as its generic vector argument. The symbol ϵ_A is a fixed positive numerical floor, and $u_\ell(x)$ is the baseline nonlinear preactivation. Finally, $\widetilde{a}_\phi^{(\ell)}$ is the scale-matched geometry modulation and α_ℓ is the multiplicative adapter scale; the design uses a shared constant α_A across depth. These quantities are

$$r_{\ell,b}(x) = H_{\ell,b}e_b(x) + U_{\ell,b}V_{\ell,b}^\top e_b(x), \tag{34}$$

$$U_{\ell,b} \in \mathbb{R}^{p_\ell \times r_A}, \qquad V_{\ell,b} \in \mathbb{R}^{d_A \times r_A}, \tag{35}$$

$$a_\ell^{\mathrm{geo}}(x) = \sqrt{\pi_{\phi,R}(x)}\, r_{\ell,R}(x) + \sqrt{\pi_{\phi,\perp}(x)}\, r_{\ell,\perp}(x), \tag{36}$$

$$\mathrm{RMS}(w) = \frac{\|w\|_2}{\sqrt{\dim(w)}}, \qquad \epsilon_A = 10^{-12}, \qquad \alpha_\ell = \alpha_A \ \ (\ell = 0, \dots, L-1), \tag{37}$$

$$\widetilde{a}_\phi^{(\ell)}(x) = \rho_\phi(x)\sqrt{\beta_{\phi,\ell}(x)}\frac{a_\ell^{\mathrm{geo}}(x)}{\mathrm{RMS}(a_\ell^{\mathrm{geo}}(x)) + \epsilon_A}\,\mathrm{RMS}(u_\ell(x)). \tag{38}$$

The fixed highway ensures that every layer receives the shared geometry code even when the low-rank residual is small. The trainable residual factorization follows the same parameter-efficiency principle as low-rank adaptation (Hu et al., 2022), while the overall compositional structure is closer to neural adapters (Houlsby et al., 2019). RMS matching keeps the geometry perturbation commensurate with the scale of the layer it modulates and avoids introducing a separate activation-specific amplitude convention.

The adapter adds trainable capacity; the exact pairing and capacity context for the end-to-end comparison are stated in Appendix G.

The training objective combines prediction fidelity with the two NRGD-supervised control targets. Let p and q denote generic two-component probability vectors and let $H^2(p, q)$ be their squared Hellinger distance. Let $\mathcal{D}_{\mathrm{train}}$ be the supervised training set, $\mathcal{D}_\rho$ the geometry centers with resolved intensity targets, and $\mathcal{D}_\pi$ the geometry centers with resolved allocation targets. The model prediction is $\widehat{G}_{\theta,\phi}$, where θ denotes backbone parameters and ϕ adapter parameters. The losses $\mathcal{L}_{\mathrm{pred}}$, $\mathcal{L}_{\mathrm{int}}$, and $\mathcal{L}_{\mathrm{alloc}}$ denote prediction, intensity-supervision, and allocation-supervision losses, and $\lambda_\rho, \lambda_\pi \geq 0$ are their fixed weights:

$$H^2(p, q) = \frac{1}{2}\sum_{k=1}^{2}(\sqrt{p_k} - \sqrt{q_k})^2, \tag{39}$$

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{|\mathcal{D}_{\mathrm{train}}|}\sum_{x \in \mathcal{D}_{\mathrm{train}}}\frac{1}{m}\|\widehat{G}_{\theta,\phi}(x) - G(x)\|_2^2, \tag{40}$$

$$\mathcal{L}_{\mathrm{int}} = \frac{1}{|\mathcal{D}_\rho|}\sum_{x \in \mathcal{D}_\rho}(\rho_\phi(x) - \rho_G^\star(x))^2, \tag{41}$$

$$\mathcal{L}_{\mathrm{alloc}} = \frac{1}{|\mathcal{D}_\pi|}\sum_{x \in \mathcal{D}_\pi} H^2(\boldsymbol{\pi}_\phi(x), \boldsymbol{\pi}_G^\star(x)), \tag{42}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda_\rho\mathcal{L}_{\mathrm{int}} + \lambda_\pi\mathcal{L}_{\mathrm{alloc}}. \tag{43}$$

The candidate weights, learning rate, adapter scale, and depth-softmax temperature are selected only in the nested development protocol. Once selected, they are frozen before held-out test evaluation. If an NRGD target family is unresolved, the corresponding geometry-supervised experiment is reported unresolved rather than silently dropping that loss term.

B Additional NRGD Experimental Analysis

This appendix expands the independent Nonlinear Response Geometry Decomposition (NRGD) study reported in the main paper. Its role is experimental rather than methodological: the main text already defines the finite-scale response Δ_G^h and r_G^h, the projected response operator $\widehat{M}_G$, the response rank R, the Dominant Nonlinear-Response Subspace (DRS) projector P_R, the local response intensity $\rho_G^\star$, and the dominant/complement allocation $\boldsymbol{\pi}_G^\star$. We therefore do not repeat those definitions here. Instead, we use the independent five-system study to examine four questions that are not visible in the compact main-paper table: finite-scale stability, sensitivity to finite geometry budgets, state-dependent local response structure, and the effect of richer input information on NRGD descriptor recovery.

All numerical values and figure assets below are unchanged from the independent study. To keep the relationship to the current method explicit, quantities introduced only for these supplementary diagnostics carry the superscript diag. They refine the same finite-scale response field for analysis, but they are not additional runtime inputs to GANR.

B.1 Diagnostic Quantities and Experimental Settings

The independent study uses Synthetic, IEEE-118, Duffing, Allen–Cahn, and shallow-water. For a geometry center x_i, unit direction n_{ik}, and finite scale h, the offline analysis computes

$$\Delta_{ik}^h := \Delta_G^h(x_i, n_{ik}) = \frac{J_G(x_i + hn_{ik}) - J_G(x_i - hn_{ik})}{2h}. \tag{44}$$

The independent five-system allocation-only study uses these full Jacobian differences to construct its diagnostic targets and final interface benchmark, as summarized in Appendix Table 8. The end-to-end GA-NAR construction described in the main method remains matrix-free and estimates directional response magnitudes from Jacobian–vector products and Hutchinson probes.

For several local analyses, let $\mathcal{V}_i$ be the valid direction set at center i. We form the diagnostic directional second moment and its total response

$$B_i^{\mathrm{diag}} = \frac{1}{|\mathcal{V}_i|} \sum_{k \in \mathcal{V}_i} r_G^h(x_i, n_{ik})\, n_{ik} n_{ik}^\top, \qquad e_i^{\mathrm{diag}} = \mathrm{tr}(B_i^{\mathrm{diag}}). \tag{45}$$

Let u_j^{diag} denote the retained directions of this independent diagnostic and let R_{diag} be its retained rank. The response carried by each retained direction and by the residual complement is

$$\begin{aligned} b_{ij}^{\mathrm{diag}} &= \max\Big((u_j^{\mathrm{diag}})^\top B_i^{\mathrm{diag}} u_j^{\mathrm{diag}}, 0 \Big), \\ b_{i,\perp}^{\mathrm{diag}} &= \max\left(e_i^{\mathrm{diag}} - \sum_{j=1}^{R_{\mathrm{diag}}} b_{ij}^{\mathrm{diag}}, 0 \right). \end{aligned} \tag{46}$$

These direction-resolved quantities are used only to expose the structure hidden by a two-component summary. Let $\epsilon = 10^{-12}$, let $\mathbf{1}$ denote the all-ones vector of length $R_{\mathrm{diag}} + 1$, and let $\bar{e}^{\mathrm{diag}}$ be the training-population mean of e_i^{diag}. The corresponding diagnostic fraction vector, bounded total-response scalar, and complete allocation vector are

$$\begin{aligned} f_i^{\mathrm{diag}} &= \frac{(b_{i1}^{\mathrm{diag}}, \ldots, b_{iR_{\mathrm{diag}}}^{\mathrm{diag}}, b_{i,\perp}^{\mathrm{diag}}) + \epsilon \mathbf{1}}{e_i^{\mathrm{diag}} + (R_{\mathrm{diag}} + 1)\epsilon}, \\ \alpha_i^{\mathrm{diag}} &= \frac{e_i^{\mathrm{diag}}}{e_i^{\mathrm{diag}} + \bar{e}^{\mathrm{diag}}}, \qquad w_i^{\mathrm{diag}} = \big(1 - \alpha_i^{\mathrm{diag}},\ \alpha_i^{\mathrm{diag}} f_i^{\mathrm{diag}}\big), \end{aligned} \tag{47}$$

Table 4: Settings of the independent five-system NRGD study. m and d are state and standardized-input dimensions. The Centers column gives the number of geometry centers, K_{dir} is the number of explicit directions per center, Mass is the spectral mass captured by R_{diag}, and $\bar{\alpha}^{\mathrm{diag}}$ is the mean bounded total-response scalar used in the interface-fidelity experiment.

System	m	d	Centers	K_{dir}	R_{diag}	R_{90}^{diag}	Mass	$\bar{\alpha}^{\mathrm{diag}}$
Synthetic	50	50	1024	12	5	5	1.000	0.469
IEEE-118	180	36	1024	12	6	5	0.951	0.469
Duffing	16	16	1024	16	12	12	0.913	0.496
Allen–Cahn	64	64	1536	12	14	23	0.769	0.471
Shallow-water	64	64	1280	16	16	44	0.568	0.491

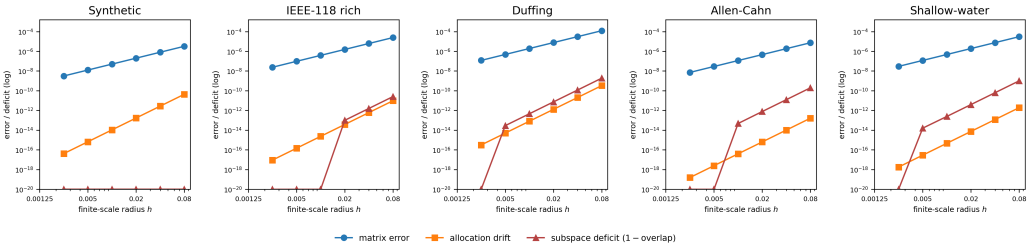


Figure 2: Finite-scale stability of the five-system NRGD diagnostic. Columns show Synthetic, IEEE-118, Duffing, Allen–Cahn, and shallow-water. Matrix discrepancy and allocation drift use the logarithmic left axis; dominant-subspace deficit $1 - \mathrm{overlap}$ uses the logarithmic right axis. The same center and direction banks are reused across radii.

Up to the common dimension factor used in the main NRGD estimator, α_i^{diag} has the same bounded relative-intensity normalization as $\rho_G^\star$. By contrast, f_i^{diag} resolves retained directions individually and is therefore a finer, basis-dependent evaluation coordinate than the main two-component DRS/complement target $\boldsymbol{\pi}_G^\star$. The scalar α_i^{diag} is therefore evaluated as a total-response descriptor, not as physical energy or intrinsic neural capacity.

Table 4 records the dimensions and geometry budgets of the five-system study. To avoid collision with the current NRGD rank R, R_{diag} denotes the retained rank used in this independent analysis and R_{90}^{diag} denotes its uncapped 90% spectral-rank diagnostic.

The contrast between R_{diag} and R_{90}^{diag} already shows that the global response spectrum is strongly system dependent. In particular, Allen–Cahn and shallow-water retain longer spectral tails than Synthetic, IEEE-118, and Duffing under this diagnostic. This motivates preserving an explicit complement channel rather than interpreting the retained directions as an exhaustive response basis.

B.2 Finite-Scale Stability and Global Response Structure

NRGD is defined at finite perturbation scale, so the first audit tests whether the measured response geometry is dominated by a particular radius choice. The same center and direction banks are reused at

$$h \in \{0.00125, 0.0025, 0.005, 0.01, 0.02, 0.04, 0.08\}. \tag{48}$$

The recorded invalid-pair rate is zero for all five systems. Figure 2 compares adjacent radii using the normalized change of the diagnostic global response matrix, drift of w_i^{diag}, and dominant-subspace deficit $1 - \mathrm{overlap}$. Because the geometry samples are fixed across radii, the curves isolate scale sensitivity rather than resampling variability.

Across the evaluated grid, all three diagnostics remain small. The largest adjacent matrix discrepancy is 1.26×10^{-4}, the minimum subspace overlap is 0.9999999977, and the maximum allocation drift is 3.52×10^{-10}; all three extrema occur for Duffing. These measurements support finite-scale stability over the evaluated range. They do not replace the scale-selection and resolution tests of the main NRGD construction protocol.

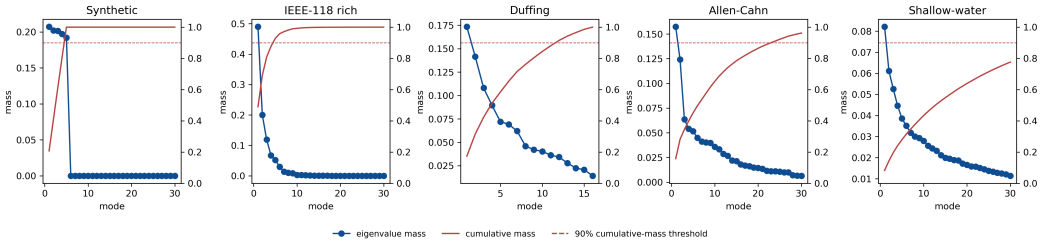


Figure 3: Global spectral concentration of the full-matrix response diagnostic at $h = 0.02$. Columns show Synthetic, IEEE-118, Duffing, Allen–Cahn, and shallow-water. Points show descending eigenvalue mass and curves show cumulative mass; the dashed line marks 90% cumulative mass and defines R_{90}^{diag} for this diagnostic.

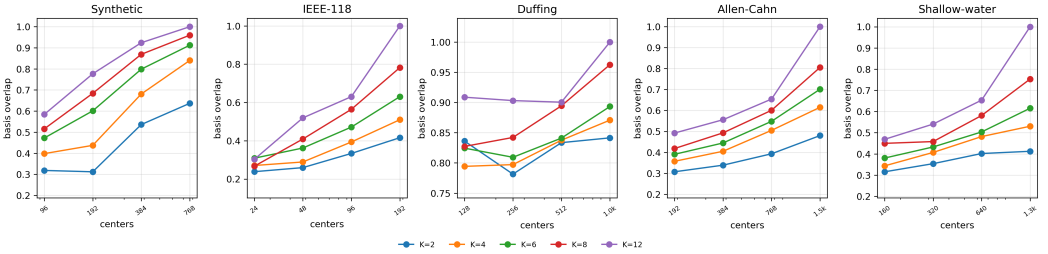


Figure 4: Subspace stability under finite center and direction budgets. The vertical axis is overlap $\|U_r^\top V_r\|_F^2/r$ with the joint maximum-budget reference, using $r = \min(r_U, r_V)$. Each curve fixes a direction budget. The unit-overlap endpoint at the joint maximum budget is true by construction; the informative behavior is the approach to that endpoint.

Figure 3 next examines global spectral concentration at $h = 0.02$. The plotted object is the trace-normalized mean of $(\Delta_{ik}^h)^\top \Delta_{ik}^h$ across the recorded center–direction pairs. It is a full-matrix diagnostic of the same finite-scale response field, not the projected response operator $\widehat{M}_G$ used by the main NRGD construction.

Synthetic concentrates essentially all spectral mass in five directions, and IEEE-118 also crosses 90% at five. Duffing reaches 90% at rank 12, whereas Allen–Cahn and shallow-water require ranks 23 and 44. Their fixed diagnostic ranks $(14, 16)$ consequently capture 76.25% and 56.91% of the mass, respectively. The spectra therefore support the central NRGD premise that response geometry differs substantially across physical systems, while making no claim that spectral rank is itself a measure of approximation difficulty or intrinsic physical dimension.

B.3 Sensitivity to Finite Geometry Budgets

The next audit varies the number of geometry centers and explicit directions using nested prefixes of fixed maximum arrays. This construction reduces unrelated resampling variability and directly tests how the estimated response subspace changes as measurement coverage increases.

Figure 4 compares each budget-limited retained subspace with the joint maximum-center, maximum-direction reference. The direction budgets are $K_{\mathrm{dir}} = (2, 4, 6, 8, 12)$ for Synthetic, IEEE-118, and Allen–Cahn and $K_{\mathrm{dir}} = (2, 4, 6, 10, 16)$ for Duffing and shallow-water. The center prefixes are $(96, 192, 384, 768)$, $(24, 48, 96, 192)$, $(128, 256, 512, 1024)$, $(192, 384, 768, 1536)$, and $(160, 320, 640, 1280)$ in system order.

Overlap generally increases with additional centers, although the trend is not strictly monotone. At the largest direction budget, Duffing is already close to its reference at the smallest center prefix, while Synthetic approaches the reference more gradually. IEEE-118, Allen–Cahn, and shallow-water require broader center coverage. Lower direction budgets remain separated from the reference even

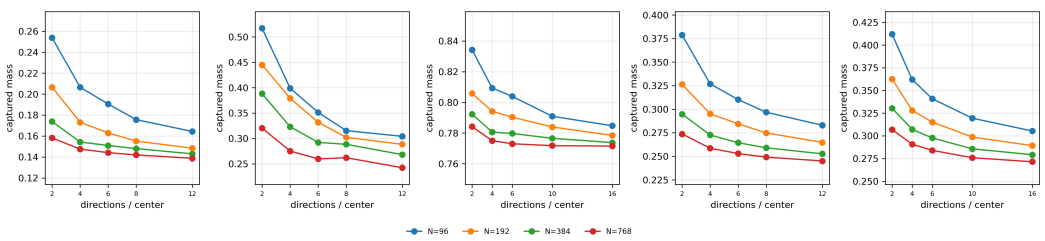


Figure 5: Spectral mass captured by the fixed diagnostic rank under finite direction budgets at $h = 0.02$. Columns show Synthetic, IEEE-118, Duffing, Allen–Cahn, and shallow-water; colored curves fix center prefixes. The plotted rank is R_{diag} from Table 4, not the main NRGD rank-selection curve.

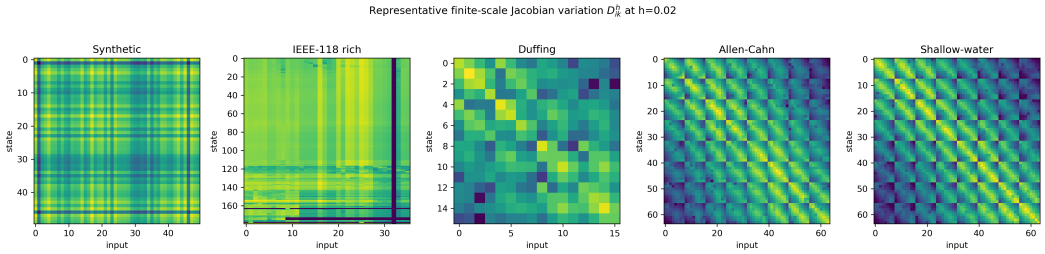


Figure 6: Representative finite-scale Jacobian variations $\Delta_G^h(x_i, n_{ik})$ at $h = 0.02$. Panels show Synthetic, IEEE-118, Duffing, Allen–Cahn, and shallow-water. State coordinates are rows and standardized-input coordinates are columns. Colors show $\log_{10}(|\Delta_G^h| + 10^{-12})$ with independent panel normalization.

at large center counts, indicating that both operating-point coverage and directional probing affect the estimated subspace.

Figure 5 complements this comparison by tracking the spectral mass captured by the fixed R_{diag} as the direction budget changes. The fixed ranks are $(5, 6, 12, 14, 16)$ in system order.

Captured mass generally decreases as additional directions expose response structure that was not visible at smaller budgets, with the largest changes concentrated at the smallest budgets. Duffing and Synthetic flatten most clearly at larger budgets, whereas IEEE-118, Allen–Cahn, and shallow-water retain stronger direction-budget dependence. This behavior supports the use of explicit geometry-resolution checks rather than treating a single finite budget as exact.

B.4 State-Dependent Local Response Structure

The global analyses above summarize magnitude, concentration, and subspace stability but do not show where the finite-scale Jacobian variation is expressed across state and input coordinates. Figure 6 therefore visualizes representative $\Delta_G^h(x_i, n_{ik})$ matrices at $h = 0.02$.

The displayed center indices are $(26, 88, 187, 162, 146)$ in system order, with direction index zero. The patterns are visibly nonuniform across both systems and coordinates: Synthetic is comparatively dispersed, IEEE-118 has pronounced row/column structure, and the remaining systems contain more localized diagonal or near-diagonal patterns. The scalar score r_G^h summarizes response magnitude and therefore cannot retain this entrywise organization; the heatmaps illustrate the structure from which the scalar directional responses are derived. Because each panel uses an independent color range, the figure should not be used for cross-system magnitude comparison.

Figure 7 examines operating-point dependence through B_i^{diag}. The first three rows show centers selected at low, median, and high e_i^{diag}, while the bottom row shows the fine-grained diagnostic fractions f_i^{diag} across retained directions and complement.

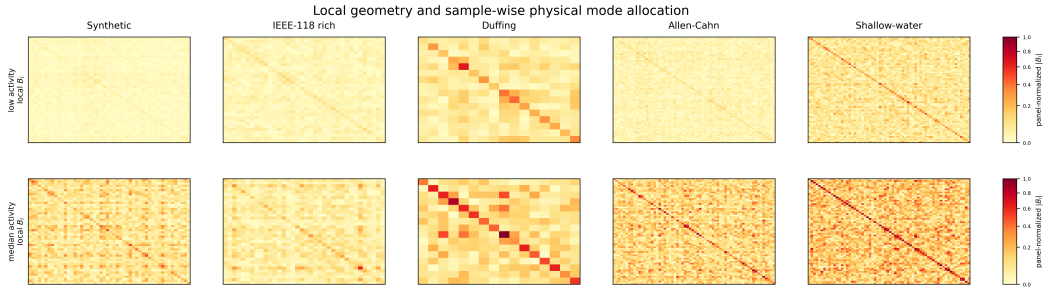


Figure 7: Local response structure at $h = 0.02$. Columns show Synthetic, IEEE-118, Duffing, Allen–Cahn, and shallow-water. The first three rows visualize B_i^{diag} at low, median, and high response activity using a shared scale within each system. The bottom row shows the fine-grained diagnostic fractions over retained directions and complement, with centers ordered by increasing total response.

Table 5: Information-rich target-recovery diagnostic on the five independent NRGD systems. Values are validation means over three training seeds. Allocation MSE averages all entries of w_i^{diag}; fraction MSE averages the entries of f_i^{diag}; fraction NMSE normalizes by the mean validation variance of the fraction channels.

System	Allocation MSE	Fraction MSE	Fraction NMSE ↓
Synthetic	1.24×10^{-5}	6.22×10^{-4}	0.673
IEEE-118	4.99×10^{-6}	1.71×10^{-5}	0.042
Duffing	5.63×10^{-7}	2.33×10^{-6}	0.004
Allen–Cahn	1.79×10^{-6}	8.23×10^{-6}	0.092
Shallow-water	1.28×10^{-6}	5.22×10^{-6}	0.086

Within each system, response magnitude and off-diagonal structure change with the operating point, and the bottom row shows corresponding variation in directional composition. This is direct evidence that a single global spectrum is insufficient to describe all local operating states. It motivates the state-dependent intensity and allocation descriptors used by NRGD, without implying that every individual basis-resolved fraction must be recovered at inference time.

B.5 Extended Interface-Fidelity Analysis

The compact main-paper table reports complete allocation MSE, total-response MSE, and fraction NMSE. Here we retain two complementary diagnostics from the same five-system experiment: an information-rich target-recovery test and a full input-only-versus-directional-sketch comparison with total-response R^2 and fraction correlation. These are descriptor-fidelity results, not end-to-end state-prediction results.

Information-rich target-recovery diagnostic. A learned descriptor predictor is first supplied with the local quantities from which the fine-grained diagnostic target is constructed,

$$\left(b_{i1}^{\mathrm{diag}}, \ldots, b_{iR_{\mathrm{diag}}}^{\mathrm{diag}}, b_{i,\perp}^{\mathrm{diag}}, e_i^{\mathrm{diag}}\right), \tag{49}$$

with feature standardization fitted on training rows only. This intentionally information-rich setting tests whether the target map itself is learnable from finite samples; it is neither an inference-time input setting nor an analytic ceiling.

All five fraction NMSE values are below one, ranging from 0.004 to 0.673. Thus, when the construction variables are exposed, the fine-grained allocation contains recoverable sample-dependent structure relative to its validation variance. The result does not imply comparable recovery from input-only features.

Input-only and directional-sketch settings. The input-only NRGD descriptor predictor uses standardized input x together with frozen-subspace features derived from the retained diagnostic

basis: signed, absolute, squared, and eigenvalue-weighted retained coordinates plus complement-norm features. The raw-input control uses x alone; the random-projector control substitutes a seeded random orthonormal basis; the intensity-only control predicts α_i^{diag} but assigns uniform fractions; and the training-mean control ignores the input.

For the richer directional-sketch setting, let $q_k \in \mathbb{R}^d$ denote the kth fixed normalized Gaussian probe and let $K = \min(4R_{\mathrm{diag}}, d)$ be the number of probes. We denote the resulting scalar response measurement by s_{ik} and its dimension-corrected trace estimate by $\widehat{e}_i^{\mathrm{diag}}$. They are

$$s_{ik} = q_k^\top B_i^{\mathrm{diag}} q_k, \qquad \widehat{e}_i^{\mathrm{diag}} = \frac{d}{K} \sum_{k=1}^{K} s_{ik}, \tag{50}$$

These measurements are computed offline from precomputed local response matrices. The learned models do not receive the exact retained-direction strengths, complement strength, or exact target trace. This is why the main paper treats the directional sketch as an information-rich diagnostic rather than an inference-time baseline.

For the expanded table, allocation MSE averages squared error over all entries of w_i^{diag}. Let V_0 denote the common validation partition and let MSE_α be the validation MSE of the bounded total-response scalar. Total-response R^2 is

$$R^2_{\mathrm{resp}} = 1 - \frac{\mathrm{MSE}_\alpha}{\max\{\mathrm{Var}_{V_0}(\alpha^{\mathrm{diag}}), 10^{-12}\}}. \tag{51}$$

Fraction NMSE divides fraction MSE by the mean per-channel target variance, and fraction correlation averages channelwise Pearson correlations with numerical denominator floors. The experiment uses an 80/20 train/validation split and training seeds $(0, 1, 2)$; reported scores are validation results after early stopping.

Table 6 shows that input-only NRGD has lower allocation MSE than the listed input-only controls on all five systems, matching the compact main-paper result. Adding the directional sketch lowers NRGD allocation MSE on every system. The improvement is not uniform across every fraction metric—for Synthetic, for example, fraction NMSE changes from 0.890 to 1.000—which reinforces the distinction between recovering aggregate response allocation and recovering each fine-grained directional fraction. The intensity-only control provides the complementary observation: accurate total-response information does not by itself recover where that response is distributed.

B.6 Summary and Scope

Together, the supplementary experiments establish four properties of the independent NRGD study. The measured finite-scale response is stable across the audited radius grid; the dominant response summary becomes more stable as center and direction coverage increases; global spectral concentration and local response structure vary substantially across systems and operating states; and the NRGD interface contains learnable structure whose recovery improves when richer local-response measurements are available. These observations support the system-side role assigned to NRGD in GA-NAR, while the end-to-end predictive claims remain those evaluated separately for the complete GA-NAR method. As throughout the main paper, the conclusions are finite-scale and protocol-defined and should not be read as claims of intrinsic curvature, exact Hessian recovery, globally unique physical modes, or universal measurement-budget convergence.

C Independent Evidence for GANR Geometry Conditioning

Appendix B asks whether the physical response geometry extracted by NRGD is stable, resolvable, and faithfully represented under finite geometry budgets. This appendix asks the complementary model-side question: is that response structure informative enough to justify geometry-conditioned nonlinear computation in GANR? The evidence is organized around two requirements. First, finite-response structure should distinguish operating regimes in a way that is relevant to predictive difficulty. Second, a response-aware nonlinear representation should be able to exploit that signal under controlled model interfaces. The independent experiments below address these requirements without redefining the current NRGD descriptors or the final GANR adapter.

Table 6: Extended NRGD interface-fidelity results under the input-only and directional-sketch settings. Values are mean±standard deviation over three training seeds. Allocation and total-response MSE are in units of 10^{-3}; fraction NMSE is dimensionless. The directional sketch provides additional finite local-response measurements but not the exact retained-direction strengths, complement strength, or exact target trace. Within each system and metric column, the original best and second-best markings are retained. The two training-mean rows are identical by construction.

System	Input setting / method	allocation MSE ↓	total-response MSE ↓	total-response R^2 ↑	fraction NMSE ↓	fraction corr. ↑
Synthetic	NRGD (input-only)	1.901 ± 0.014	7.707 ± 0.090	0.595 ± 0.005	$\mathbf{0.890 \pm 0.037}$	0.081 ± 0.017
	NRGD + directional sketch	$\mathbf{0.241 \pm 0.002}$	0.736 ± 0.004	0.961 ± 0.000	1.000 ± 0.015	$\mathbf{0.101 \pm 0.014}$
	Raw-input control (input-only)	4.848 ± 0.056	19.735 ± 0.393	-0.038 ± 0.021	2.507 ± 0.414	-0.004 ± 0.021
	Raw-input control + sketch	0.275 ± 0.008	0.852 ± 0.005	0.955 ± 0.000	1.019 ± 0.019	0.050 ± 0.025
	Random-projector control (input-only)	4.781 ± 0.141	18.645 ± 0.196	0.019 ± 0.010	2.720 ± 0.253	-0.013 ± 0.010
	Random-projector control + sketch	0.276 ± 0.007	0.836 ± 0.012	0.956 ± 0.001	1.055 ± 0.018	0.011 ± 0.016
	Intensity-only control (input-only)	22.995 ± 0.356	18.515 ± 1.245	0.026 ± 0.066	319.281 ± 15.640	0.007 ± 0.006
	Intensity-only control + sketch	20.414 ± 0.000	$\mathbf{0.730 \pm 0.001}$	$\mathbf{0.962 \pm 0.000}$	105.897 ± 0.000	0.020 ± 0.029
	Training-mean control (input-only)	4.723 ± 0.000	19.054 ± 0.000	-0.003 ± 0.000	1.006 ± 0.000	0.000 ± 0.000
	Training-mean control + sketch	4.723 ± 0.000	19.054 ± 0.000	-0.003 ± 0.000	1.006 ± 0.000	0.000 ± 0.000
IEEE-118	NRGD (input-only)	0.669 ± 0.011	3.355 ± 0.042	0.682 ± 0.004	1.068 ± 0.040	-0.026 ± 0.005
	NRGD + directional sketch	$\mathbf{0.168 \pm 0.000}$	0.487 ± 0.003	0.954 ± 0.000	$\mathbf{0.952 \pm 0.004}$	$\mathbf{0.145 \pm 0.016}$
	Raw-input control (input-only)	0.908 ± 0.020	4.044 ± 0.168	0.617 ± 0.016	2.218 ± 0.157	0.026 ± 0.023
	Raw-input control + sketch	0.176 ± 0.004	0.510 ± 0.003	0.952 ± 0.000	0.979 ± 0.030	0.089 ± 0.010
	Random-projector control (input-only)	0.880 ± 0.008	4.358 ± 0.032	0.587 ± 0.003	1.254 ± 0.056	-0.009 ± 0.017
	Random-projector control + sketch	0.205 ± 0.013	0.552 ± 0.057	0.948 ± 0.005	1.108 ± 0.051	0.064 ± 0.026
	Intensity-only control (input-only)	14.914 ± 0.037	3.800 ± 0.177	0.640 ± 0.017	155.312 ± 14.930	-0.003 ± 0.012
	Intensity-only control + sketch	14.147 ± 0.000	$\mathbf{0.471 \pm 0.001}$	$\mathbf{0.955 \pm 0.000}$	175.048 ± 0.000	-0.003 ± 0.011
	Training-mean control (input-only)	2.157 ± 0.000	10.577 ± 0.000	-0.002 ± 0.000	1.008 ± 0.000	0.000 ± 0.000
	Training-mean control + sketch	2.157 ± 0.000	10.577 ± 0.000	-0.002 ± 0.000	1.008 ± 0.000	0.000 ± 0.000
Duffing	NRGD (input-only)	0.248 ± 0.003	1.544 ± 0.034	0.131 ± 0.019	1.008 ± 0.002	0.013 ± 0.002
	NRGD + directional sketch	$\mathbf{0.132 \pm 0.000}$	0.281 ± 0.005	0.842 ± 0.003	$\mathbf{0.855 \pm 0.000}$	$\mathbf{0.383 \pm 0.003}$
	Raw-input control (input-only)	0.267 ± 0.004	1.597 ± 0.022	0.100 ± 0.012	1.116 ± 0.014	-0.016 ± 0.017
	Raw-input control + sketch	0.134 ± 0.001	0.284 ± 0.006	0.840 ± 0.004	0.859 ± 0.006	0.374 ± 0.004
	Random-projector control (input-only)	0.264 ± 0.011	1.624 ± 0.082	0.085 ± 0.046	1.087 ± 0.036	-0.003 ± 0.012
	Random-projector control + sketch	0.153 ± 0.001	0.275 ± 0.007	0.845 ± 0.004	1.016 ± 0.004	0.064 ± 0.004
	Intensity-only control (input-only)	0.720 ± 0.008	1.839 ± 0.230	-0.036 ± 0.129	4.892 ± 0.249	0.000 ± 0.000
	Intensity-only control + sketch	0.615 ± 0.000	$\mathbf{0.265 \pm 0.000}$	$\mathbf{0.851 \pm 0.000}$	4.685 ± 0.000	-0.000 ± 0.002
	Training-mean control (input-only)	0.266 ± 0.000	1.776 ± 0.000	-0.000 ± 0.000	1.004 ± 0.000	0.000 ± 0.000
	Training-mean control + sketch	0.266 ± 0.000	1.776 ± 0.000	-0.000 ± 0.000	1.004 ± 0.000	0.000 ± 0.000
Allen–Cahn	NRGD (input-only)	0.079 ± 0.001	0.776 ± 0.011	0.948 ± 0.001	1.074 ± 0.039	-0.006 ± 0.003
	NRGD + directional sketch	$\mathbf{0.033 \pm 0.000}$	0.150 ± 0.003	0.990 ± 0.000	$\mathbf{0.993 \pm 0.003}$	$\mathbf{0.055 \pm 0.007}$
	Raw-input control (input-only)	0.109 ± 0.002	1.056 ± 0.029	0.929 ± 0.002	1.800 ± 0.235	-0.001 ± 0.003
	Raw-input control + sketch	0.033 ± 0.000	0.159 ± 0.004	0.989 ± 0.000	1.004 ± 0.014	0.051 ± 0.007
	Random-projector control (input-only)	0.115 ± 0.009	1.109 ± 0.081	0.926 ± 0.005	1.240 ± 0.047	-0.018 ± 0.022
	Random-projector control + sketch	0.043 ± 0.002	0.169 ± 0.008	0.989 ± 0.001	1.319 ± 0.038	0.010 ± 0.013
	Intensity-only control (input-only)	7.740 ± 0.168	1.135 ± 0.051	0.924 ± 0.003	398.857 ± 7.798	-0.001 ± 0.004
	Intensity-only control + sketch	7.593 ± 0.000	$\mathbf{0.140 \pm 0.000}$	$\mathbf{0.991 \pm 0.000}$	391.275 ± 0.000	-0.001 ± 0.001
	Training-mean control (input-only)	1.498 ± 0.000	15.007 ± 0.000	-0.003 ± 0.000	1.002 ± 0.000	0.000 ± 0.000
	Training-mean control + sketch	1.498 ± 0.000	15.007 ± 0.000	-0.003 ± 0.000	1.002 ± 0.000	0.000 ± 0.000
Shallow-water	NRGD (input-only)	0.030 ± 0.000	0.278 ± 0.005	0.947 ± 0.001	1.080 ± 0.020	0.015 ± 0.014
	NRGD + directional sketch	$\mathbf{0.021 \pm 0.000}$	0.110 ± 0.001	0.979 ± 0.000	$\mathbf{0.927 \pm 0.001}$	$\mathbf{0.117 \pm 0.004}$
	Raw-input control (input-only)	0.042 ± 0.002	0.328 ± 0.015	0.937 ± 0.003	1.754 ± 0.024	0.000 ± 0.007
	Raw-input control + sketch	0.022 ± 0.000	0.121 ± 0.004	0.977 ± 0.001	0.940 ± 0.007	0.085 ± 0.007
	Random-projector control (input-only)	0.038 ± 0.002	0.331 ± 0.032	0.937 ± 0.006	1.309 ± 0.026	-0.006 ± 0.005
	Random-projector control + sketch	0.032 ± 0.001	0.138 ± 0.003	0.974 ± 0.001	1.602 ± 0.034	0.038 ± 0.011
	Intensity-only control (input-only)	6.774 ± 0.028	0.332 ± 0.022	0.936 ± 0.004	503.327 ± 22.674	-0.001 ± 0.002
	Intensity-only control + sketch	6.663 ± 0.000	$\mathbf{0.108 \pm 0.000}$	$\mathbf{0.979 \pm 0.000}$	478.460 ± 0.000	-0.002 ± 0.005
	Training-mean control (input-only)	0.447 ± 0.000	5.233 ± 0.000	-0.002 ± 0.000	1.002 ± 0.000	-0.000 ± 0.000
	Training-mean control + sketch	0.447 ± 0.000	5.233 ± 0.000	-0.002 ± 0.000	1.002 ± 0.000	-0.000 ± 0.000

C.1 Geometry Interface and Evidence Boundary

All notation follows the main manuscript. The physical parameter and state are q and z, their train-standardized coordinates are $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^m$, and the standardized parameter-to-solution map is $G : x \mapsto y$. For a unit direction n and finite scale h, the physical nonlinear response is described by the centered Jacobian variation $\Delta_G^h(x, n)$ and squared response magnitude $r_G^h(x, n)$ from Equation 4. Current NRGD supplies the Dominant Nonlinear-Response Subspace projector P_R, local nonlinear-response intensity $\rho_G^\star(x)$, and dominant/complement allocation $\boldsymbol{\pi}_G^\star(x)$. In GANR, P_R defines the dual-subspace decomposition, while $\rho_G^\star$ and $\boldsymbol{\pi}_G^\star$ supervise the corresponding controller outputs; the learned depth allocation remains task-driven. At the level of the training interface, the information flow is therefore

$$\big(P_R, \rho_G^\star(x), \boldsymbol{\pi}_G^\star(x)\big) \xrightarrow[\text{geometry supervision}]{\text{GANR}} \big(\rho_\phi(x), \boldsymbol{\pi}_\phi(x), \boldsymbol{\beta}_\phi(x)\big) \longrightarrow \widetilde{a}_\phi^{(\ell)}(x), \tag{52}$$

with the detailed controller and modulation definitions given in Section 4 and Appendix A.3.

This independent study uses an earlier scalar finite-response proxy for stratification and an earlier response-aware GANR formulation. To keep the evidence boundary explicit, the scalar proxy is not identified with $\rho_G^\star(x)$, and the auxiliary GANR formulation is not identified with the final dual-subspace, layer-wise adapter. These experiments test the representation relevance of finite-scale response information: whether response regimes are associated with different predictive behavior and whether conditioning nonlinear computation on response information is useful under matched interfaces. Claims about the exact controller targets, layer-wise modulation, and complete GA-NAR construction rely on the experiments defined in the main manuscript.

C.2 Controlled GANR Experimental Settings

The evidence uses two complementary settings. The first contains five systems—Controlled, Duffing, IEEE-118, Allen–Cahn, and shallow-water—and compares fixed-activation models with an auxiliary response-aware GANR formulation. Controlled and Duffing use 50,000/5,000/10,000 train/validation/test samples; IEEE-118, Allen–Cahn, and shallow-water use 10,000/1,000/2,000. Geometry diagnostics use 1,024 held-out probe centers and 16 directions per center, and means and sample standard deviations use seeds 11, 29, and 47.

Some comparisons include a common equation-derived local linearization, referred to below as the shared-physics condition. This object is fixed across compared models and is distinct from the governing-equation residual used by the final GA-NAR evaluation. The second setting places the candidate nonlinear modules in a common residual-MLP interface under its own fixed data and training protocol. Numerical values are interpreted within each setting rather than ranked across settings.

The purpose of these experiments is not to repeat the NRGD resolution study in Appendix B. Instead, they provide an independent bridge from measured physical response structure to model-side nonlinear representation, which is the design premise of GANR.

C.3 Response-Conditioned Predictive Difficulty

A prerequisite for geometry-conditioned nonlinear representation is that physical response regime carries information relevant to the learning problem. The scalar finite-response proxy provides an independent way to test this premise. Test points are partitioned into five equal-frequency groups from Q1 (lowest response) to Q5 (highest response). The scalar is an earlier aggregate of finite-response information and is used only for stratification; it is not identified with the final bounded NRGD intensity target $\rho_G^\star(x)$.

Figure 8 shows that prediction error changes across finite-response strata in a system- and model-dependent manner. The evidence does not assert a universal monotonic relation between response magnitude and error. Rather, it shows that aggregate value error can hide operating-regime dependence. This is precisely the type of heterogeneity for which a uniform nonlinear treatment provides no explicit mechanism, motivating GANR to condition nonlinear computation on system-derived response structure.

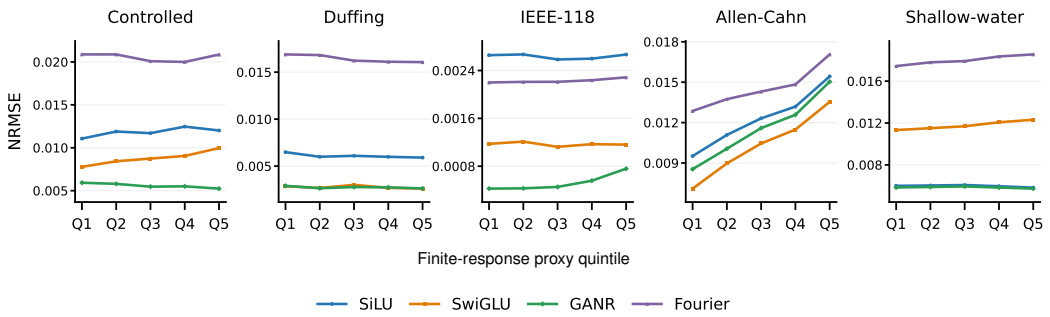


Figure 8: Auxiliary response-stratified prediction error. Panels show the five systems (Controlled, Duffing, IEEE-118, Allen–Cahn, and shallow-water) and the four compared models (SiLU, SwiGLU, GANR, and Fourier). Q1–Q5 are equal-frequency quintiles of the solver-derived finite-response proxy, and the vertical axis is NRMSE. Each point is the arithmetic mean over seeds 11, 29, and 47 under the shared-physics condition. Panels use system-specific vertical scales for readability.

C.4 Response-Aware Nonlinear Representation Under Controlled Settings

The main paper reports the matched No physics/Shared physics factorial once, in Table 2. In both conditions GANR is lowest on Controlled, Duffing, IEEE-118, and shallow-water, while Allen–Cahn remains the exception. Because the equation-derived linearization is absent from every model in one condition and shared identically by every model in the other, this comparison shows that the GANR advantage is not explained by unequal auxiliary-physics access. We do not duplicate those numerical rows here.

The additional test below changes the model interface while keeping the auxiliary information common. Its purpose is to test whether the same system-dependent pattern persists when the candidate nonlinearities are inserted into a shared residual-MLP backbone.

C.5 Strict Plug-In Test of Response-Aware Nonlinearity

A stricter plug-in setting places the compared nonlinearities in the same residual-MLP interface. The backbone construction, initialization protocol, fixed physics linearization, and geometry-loss type are held common, while model parameters are trained jointly rather than frozen. All reported rows use seeds 11, 29, and 47, and every method receives the same physics linearization. For the ordinary systems the split is 30,000/3,000/6,000; the three larger systems use 10,000/1,000/2,000 train/validation/test points, 768 held-out probe centers, and 16 directions per center.

The GANR row has lower NRMSE than the listed fixed-activation alternatives on Controlled, Duffing, IEEE-118, and shallow-water, whereas SwiGLU is lower on Allen–Cahn. Reproducing the same system-dependent pattern under a common plug-in interface strengthens the interpretation that response-aware nonlinear representation can provide useful model-side conditioning, while also showing that the benefit is not universal across all physical systems.

Table 7: Auxiliary strict plug-in benchmark under a common residual-MLP interface. All methods use the shared residual MLP, common physics linearization, and geometry loss; values are means ± sample standard deviations over seeds 11, 29, and 47, with lower NRMSE better.

System	GANR	GELU	ReLU	SiLU	SwiGLU
Controlled	**0.007194 ± 0.000044**	0.009604 ± 0.000133	0.016843 ± 0.000090	0.012629 ± 0.000140	0.008630 ± 0.000055
Duffing	**0.006111 ± 0.000061**	0.009993 ± 0.000165	0.013587 ± 0.000156	0.010998 ± 0.000138	0.007166 ± 0.000148
IEEE-118	**0.000558 ± 0.000054**	0.000668 ± 0.000044	0.000576 ± 0.000046	0.000665 ± 0.000038	0.000845 ± 0.000076
Allen–Cahn	0.012621 ± 0.000016	0.013601 ± 0.000110	0.014450 ± 0.000134	0.013716 ± 0.000107	**0.011660 ± 0.000024**
Shallow-water	**0.008026 ± 0.000047**	0.008403 ± 0.000040	0.010504 ± 0.000035	0.008074 ± 0.000052	0.014502 ± 0.000018

C.6 What the Independent Evidence Supports About GANR

Appendices B and C deliberately answer different questions. Appendix B addresses measurement validity: whether finite-scale response geometry can be resolved stably and summarized faithfully by NRGD. The present appendix addresses representation relevance: whether response regimes coincide with heterogeneous predictive behavior and whether a response-aware nonlinear representation can exploit that information under controlled auxiliary settings. Main-paper Table 2 provides the primary controlled comparison by showing the same model comparison with the auxiliary physics factor absent for all models and shared identically by all models.

These independent experiments do not provide an alternative definition of NRGD and do not constitute an isolated ablation of every component of the final GANR adapter. In particular, the scalar response proxy is not $\rho_G^\star(x)$, the auxiliary GANR row in Table 2 is not the final dual-subspace, layer-wise adapter, and the shared-physics factor is not the physical-equation residual used in the primary GA-NAR evaluation. Their role is narrower but directly relevant: they independently support the two premises on which GANR is built—that nonlinear-response structure varies across operating regimes in a prediction-relevant way, and that explicitly response-aware nonlinear computation can convert such structure into predictive gains under controlled interfaces.

Taken together, the evidence chain is separated by function: NRGD establishes measurable system-side response structure, the GANR studies test whether that structure is useful for nonlinear representation, and the GA-NAR evaluation measures the end-to-end predictive effect of the proposed method.

D Theoretical Analysis and Proofs

The analysis supports three specific design claims. Proposition 1 shows that finite value and first-derivative supervision does not identify second-order response geometry. Proposition 2 identifies the smooth-limit response as a positive-semidefinite Hessian–Gram form and derives the spherical moment used by NRGD. Proposition 3 bounds the GANR modulation scale and establishes exact recovery of the baseline when the modulation is zero. These results justify the proposed design ingredients; they are not an optimality theorem and do not imply that larger physical response provably requires larger neural nonlinear capacity. The remaining subsections provide supporting invariance details and an analytic oracle.

D.1 Proposition 1: finite-sample non-identifiability of nonlinear-response geometry

The first proposition formalizes why value fitting alone cannot identify the response geometry used by NRGD. Let $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ be any finite collection of distinct sample locations, let $f : \mathbb{R}^d \to \mathbb{R}^m$ be any smooth vector-valued map, let $e_a \in \mathbb{R}^m$ select one output coordinate, and let $c \neq 0$ be a scalar perturbation amplitude. We construct a polynomial b that vanishes together with its first derivative at every sample. For a fixed sample x_i, the auxiliary factor $c_i(x)$ contains all terms in the product except the factor associated with x_i. The perturbed map $\widetilde{f}$ is then

$$\{x_i\}_{i=1}^N \subset \mathbb{R}^d, \qquad f : \mathbb{R}^d \to \mathbb{R}^m, \qquad e_a \in \mathbb{R}^m, \qquad c \neq 0, \tag{53}$$

$$b(x) = \prod_{i=1}^{N} \|x - x_i\|_2^2, \qquad \widetilde{f}(x) = f(x) + c\, e_a b(x), \tag{54}$$

$$c_i(x) = \prod_{j \neq i} \|x - x_j\|_2^2, \qquad b(x) = \|x - x_i\|_2^2 c_i(x), \tag{55}$$

$$b(x_i) = 0, \qquad \nabla b(x_i) = 0, \qquad \nabla^2 b(x_i) = 2c_i(x_i) I_d \neq 0. \tag{56}$$

The original and perturbed maps therefore agree on all sampled values and first derivatives, yet their second-order behavior differs. Since the smooth-limit response operator depends on second derivatives, finite pointwise supervision cannot identify that operator without additional information.

D.2 Proposition 2: smooth-limit Hessian–Gram response identity

Let G_a denote output component a of the solution map, and let $H_a(x)$ denote the Hessian of that component with respect to the standardized input. For a unit input direction n, D_n denotes the directional derivative. The directional derivative of the ath Jacobian row is obtained by applying H_a to n. Summing the squared row norms over all m outputs defines the positive-semidefinite matrix $T_G(x)$:

$$D_n \nabla G_a(x)^\top = (H_a(x)n)^\top, \qquad \|D_n J_G(x)\|_F^2 = \sum_{a=1}^{m} \|H_a(x)n\|_2^2, \tag{57}$$

$$\|D_n J_G(x)\|_F^2 = n^\top T_G(x) n, \qquad T_G(x) = \sum_{a=1}^{m} H_a(x)^\top H_a(x) \succeq 0. \tag{58}$$

Twice continuous differentiability guarantees that the centered finite-scale Jacobian difference converges to this directional derivative as the perturbation scale tends to zero. The Gram structure explains why the limiting response is nonnegative in every direction and why a spectral decomposition is meaningful.

D.3 Spherical moment identity

Let S^{d-1} denote the unit sphere in $\mathbb{R}^d$, and let n be uniformly distributed on that sphere. For coordinate indices i, j, k, l, let δ_{ij} denote the Kronecker delta. Let $T \in \mathbb{R}^{d\times d}$ be any symmetric matrix. The fourth-order spherical moment and its contraction with the quadratic form $n^\top T n$ are

$$S^{d-1} = \{n \in \mathbb{R}^d : \|n\|_2 = 1\}, \qquad \delta_{ij} = \mathbf{1}[i = j], \tag{59}$$

$$\mathbb{E}[n_i n_j n_k n_l] = \frac{\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}}{d(d+2)}, \tag{60}$$

$$\mathbb{E}_n[(n^\top T n) n n^\top] = \frac{\mathrm{tr}(T) I_d + 2T}{d(d+2)}. \tag{61}$$

Taking the trace and normalizing the response moment gives the backprojection used in NRGD whenever the mean response trace is positive. The identity also shows why an isotropic response contributes an identity component that should not be mistaken for directional anisotropy.

D.4 Trace decomposition and basis invariance of the dominant/complement allocation

Let P_R be the rank-R DRS projector from the main method. Let n_R be uniformly distributed on the unit sphere inside $\mathrm{range}(P_R)$, and let $n_\perp$ be uniformly distributed on the unit sphere inside the complementary subspace $\mathrm{range}(I_d - P_R)$. The smooth-limit directional response is denoted by r_G^0. Using the response operator $T_G(x)$ defined above, the second moments and projected response traces are

$$\mathbb{E}[n_R n_R^\top] = \frac{P_R}{R}, \qquad \mathbb{E}[n_\perp n_\perp^\top] = \frac{I_d - P_R}{d - R}, \tag{62}$$

$$r_G^0(x, n) = n^\top T_G(x) n, \qquad R\,\mathbb{E}[r_G^0(x, n_R)] = \mathrm{tr}(P_R T_G(x)), \tag{63}$$

$$(d - R)\,\mathbb{E}[r_G^0(x, n_\perp)] = \mathrm{tr}((I_d - P_R) T_G(x)), \tag{64}$$

$$\mathrm{tr}(T_G) = \mathrm{tr}(P_R T_G) + \mathrm{tr}((I_d - P_R) T_G). \tag{65}$$

Only the projectors appear in these identities. Consequently, an orthogonal rotation of the basis within either subspace leaves the population allocation unchanged. The result remains relative to the standardized input metric fixed by the training data.

D.5 Proposition 3: scale control and exact baseline recovery of the layer-wise adapter

The GANR scale-matching rule controls the adapter magnitude relative to the baseline nonlinear preactivation. For a fixed layer ℓ, let a_ℓ^{geo} be the raw geometry direction, u_ℓ the baseline preactivation,

ρ_ϕ the sample-dependent intensity, $\beta_{\phi,\ell}$ the depth weight, and ϵ_A the numerical floor defined above. The scale-matched modulation $\widetilde{a}_\phi^{(\ell)}$ therefore satisfies

$$\widetilde{a}_\phi^{(\ell)} = \rho_\phi \sqrt{\beta_{\phi,\ell}} \frac{a_\ell^{\mathrm{geo}}}{\mathrm{RMS}(a_\ell^{\mathrm{geo}}) + \epsilon_A} \mathrm{RMS}(u_\ell), \tag{66}$$

$$\frac{\mathrm{RMS}(\widetilde{a}_\phi^{(\ell)})}{\mathrm{RMS}(u_\ell)} = \rho_\phi \sqrt{\beta_{\phi,\ell}} \frac{\mathrm{RMS}(a_\ell^{\mathrm{geo}})}{\mathrm{RMS}(a_\ell^{\mathrm{geo}}) + \epsilon_A}, \tag{67}$$

$$\frac{\mathrm{RMS}(\widetilde{a}_\phi^{(\ell)})}{\mathrm{RMS}(u_\ell)} \le \rho_\phi \sqrt{\beta_{\phi,\ell}}, \tag{68}$$

$$\frac{\mathrm{RMS}(\alpha_\ell \widetilde{a}_\phi^{(\ell)})}{\mathrm{RMS}(u_\ell)} \le |\alpha_\ell| \rho_\phi \sqrt{\beta_{\phi,\ell}}, \tag{69}$$

$$\sum_{\ell=0}^{L-1} \alpha_A^2 \rho_\phi^2 \beta_{\phi,\ell} = \alpha_A^2 \rho_\phi^2. \tag{70}$$

The last identity uses the shared layer scale and the simplex constraint on the depth allocation. These are scale-control statements for the explicit adapter path only. Exact baseline recovery follows when every adapter modulation is set to zero because the residual backbone and activation definition are otherwise unchanged.

D.6 Analytic quartic oracle

The analytical oracle isolates the stochastic and spectral parts of NRGD from numerical physical solvers. Let d_Q denote the oracle input dimension, let x_j be its independent coordinates, and let a_j be deterministic anisotropy coefficients increasing linearly with coordinate index j. The scalar oracle map is G_Q. Its Jacobian row is denoted by J_Q, its finite-scale directional Jacobian variation by Δ_Q^h, and its smooth-limit response operator by T_Q. The population-normalized response operator is M_Q^{pop}, the 90%-mass oracle rank is R_{90}^{oracle}, and N_Q is the number of analytical samples:

$$d_Q = 16, \qquad N_Q = 1024, \qquad R_{90}^{\mathrm{oracle}} = 13, \tag{71}$$

$$x_j \sim \mathrm{Unif}[-\sqrt{3}, \sqrt{3}], \qquad a_j = 1 + \frac{j-1}{15}, \tag{72}$$

$$G_Q(x) = \frac{1}{4} \sum_{j=1}^{16} a_j x_j^4, \qquad [J_Q(x)]_j = a_j x_j^3, \tag{73}$$

$$[\Delta_Q^h(x,n)]_j = a_j \left(3x_j^2 n_j + h^2 n_j^3\right), \tag{74}$$

$$T_Q(x) = \mathrm{diag}(9a_1^2 x_1^4, \ldots, 9a_{16}^2 x_{16}^4), \tag{75}$$

$$M_Q^{\mathrm{pop}} = \frac{\mathrm{diag}(a_1^2, \ldots, a_{16}^2)}{\sum_{j=1}^{16} a_j^2}. \tag{76}$$

Because all quantities are available in closed form, Q0 can independently validate the Hutchinson estimator, spherical backprojection, simplex projection, and rank-selection logic. No neural model is trained on this oracle.

E Benchmark Definitions

E.1 Formal five-system NRGD benchmark specifications

The formal NRGD interface study uses five solver-defined systems. The dimensions below are physical dimensions before standardization. For the grid-based systems, L denotes the five-point nearest-neighbour Laplacian. Physical solvers are used only for offline geometry construction; the learned descriptor interface itself does not invoke a solver at inference time.

Synthetic $50 \to 50$. The parameter and state are $q, z \in \mathbb{R}^{50}$ and satisfy

$$F(z, q) = z + 0.8C^{\mathsf{T}} \tanh(Cz) - q = 0, \tag{77}$$

where $C \in \mathbb{R}^{5\times 50}$ has orthonormal rows generated with seed 2025. Parameters follow $q \sim \mathcal{N}(0, I_{50})$. States are obtained by a monotone Newton solve with tolerance 10^{-12}; geometry uses the analytic full Jacobian and 12 normalized-Gaussian directions per center.

Coupled Duffing. Here $q, z \in \mathbb{R}^{16}$ are defined on a 4×4 grid and satisfy

$$F(z, q) = (I + 0.35L)z + c \odot z^{\odot 3} - q = 0, \qquad c_j = 0.85[1 + 0.35 \sin(2\pi \operatorname{col}(j)/4)]\,. \tag{78}$$

We sample $q \sim 0.9\mathcal{N}(0, I_{16})$. Newton iteration uses the analytic F_z, tolerance 10^{-10}, and at most 60 iterations; 16 normalized-Gaussian directions are used per center.

2D Allen–Cahn. The parameter and state have dimension 64 on an 8×8 grid, with steady reaction–diffusion equilibrium

$$F(z, q) = (0.16L + 0.9I)z + 0.55z^{\odot 3} - q = 0. \tag{79}$$

The parameter population uniformly selects one of four normalized spatial patterns, scales it by $\mathcal{U}(0.35, 1.25)$, and adds i.i.d. $\mathcal{N}(0, 0.04^2)$ pixel noise. Newton iteration uses analytic F_z, tolerance 10^{-10}, at most 60 iterations, and 12 normalized-Gaussian directions per center.

2D shallow-water. The parameter is $q \in \mathbb{R}^{64}$ on an 8×8 grid and the state $z \in \mathbb{R}^{64}$ is the free-surface anomaly. With reference depth $H_0 = 1$, the implicit map is

$$F(z, q) = (0.10L + 0.65I)z + 0.72z^{\odot 2} + 0.50z^{\odot 3} - q = 0. \tag{80}$$

The parameter population uniformly selects one of four normalized spatial patterns, scales it by $\mathcal{U}(0.30, 1.10)$, and adds i.i.d. $\mathcal{N}(0, 0.04^2)$ pixel noise. Newton iteration uses analytic F_z, tolerance 10^{-10}, at most 60 iterations, and 16 normalized-Gaussian directions per center.

IEEE-118 rich. The parameter $q \in \mathbb{R}^{36}$ contains 12 regional-load multipliers, 8 generator-redispatch controls, 8 voltage shifts, and 8 shunt controls. The state $z \in \mathbb{R}^{180}$ contains non-reference voltage angles and PQ-bus voltage magnitudes. The reduced AC power-flow residual is

$$F(z, q) = \big[P_{\text{calc}} - P_{\text{spec}},\, Q_{\text{calc}} - Q_{\text{spec}}\big] = 0, \tag{81}$$

with ZIP loads, redispatch, voltage setpoints, and shunts on the fixed case118 topology. Regional-load multipliers follow $\mathcal{U}(0.55, 1.45)$; redispatch and shunt controls follow $0.45\mathcal{N}(0, I_8)$; voltage shifts follow $0.10\mathcal{N}(0, I_8)$. An analytic Newton solve uses tolerance 10^{-10}, at most 50 iterations, base-branch continuation, a dense reduced Jacobian, and 12 normalized-Gaussian directions per center.

E.2 Analytical and auxiliary synthetic benchmarks

Q0: anisotropic quartic analytical oracle. Q0 is the closed-form quartic construction defined above. Its role is methodological rather than predictive: the exact response operator and response rank are known, so failures can be attributed to stochastic estimation or spectral processing rather than to a nonlinear physical solve.

Synthetic ($50 \to 50$). The independent NRGD study uses the 50-dimensional implicit benchmark defined in Appendix E.1. It differs from Q0 in purpose: Q0 provides a closed-form analytical oracle, whereas the $50 \to 50$ system tests descriptor recovery and geometry resolution for a moderately high-dimensional solver-defined map.

Controlled implicit nonlinear system. The GANR factorial study uses an eight-dimensional synthetic implicit system with parameter $q \in \mathbb{R}^8$ and state $z \in \mathbb{R}^8$. Its parameter-to-solution map $G_{\mathrm{C}} : q \mapsto z$ is defined by the unique solution of

$$F_{\mathrm{C}}(z, q) = Az + \alpha \tanh\big(\ell(Bz + Cq)\big) + Dq + \beta E\big[(Pq) \odot (Qq)\big] = 0, \tag{82}$$

where $A, B, C, D, E, P, Q \in \mathbb{R}^{8\times 8}$ are fixed matrices, $\odot$ denotes element-wise multiplication, and $\alpha = 0.4$, $\ell = 2.0$, and $\beta = 0.5$. We use

$$A = \mathrm{diag}(1, 5^{1/7}, \ldots, 5), \qquad B = AN, \qquad N = U\,\mathrm{diag}(1, 0.8, 0.6, 0.4)V^\top, \tag{83}$$

with $U, V \in \mathbb{R}^{8\times 4}$ having orthonormal columns. Hence $\alpha\ell\|A^{-1}B\|_2 = 0.8 < 1$, giving a contractive fixed-point construction and a well-defined local solution branch. With $F_{\mathrm{C},z}$ and $F_{\mathrm{C},q}$ denoting the state and parameter Jacobians of equation 82, the response Jacobian follows from the implicit-function relation

$$J_{G_\mathrm{C}}(q) = -F_{\mathrm{C},z}^{-1} F_{\mathrm{C},q}. \tag{84}$$

Training parameters are sampled from $\mathrm{Unif}([-0.99, 0.99]^8)$, validation and test parameters from $\mathrm{Unif}([-0.85, 0.85]^8)$, and states are obtained by damped Newton iteration.

P1: IEEE-118 AC power flow. The power-system benchmark uses the standard IEEE-118 test case from MATPOWER together with its steady-state AC power-flow formulation (Zimmerman et al., 2011). Let s_{load} be a global load scale and let ϵ_i be an independent bus-specific perturbation for load bus i. Their product defines the sampled load multiplier ℓ_i. The physical state z concatenates all non-slack voltage angles $\theta_{\text{non-slack}}$ and PQ-bus voltage magnitudes V_{PQ}:

$$s_{\mathrm{load}} \sim \mathrm{Unif}[0.85, 1.15], \qquad \epsilon_i \sim \mathrm{Unif}[-0.05, 0.05], \tag{85}$$

$$\ell_i = s_{\mathrm{load}}(1 + \epsilon_i), \qquad z = [\theta_{\text{non-slack}}, V_{\mathrm{PQ}}]. \tag{86}$$

The parameter vector consists of the load multipliers at buses with nonzero nominal demand. The slack bus fixes the reference angle, PV buses retain prescribed voltage magnitude, and PQ buses retain prescribed active and reactive demand under the standard AC power-flow convention (Zimmerman et al., 2011). Topology, bus types, and transformer taps are fixed, while reactive-limit switching is disabled so that residual ordering and state dimension remain constant over the selected solution branch. Labels and geometry endpoints are obtained by deterministic continuation from the reference operating point.

P2: Kuramoto-64. The synchronization benchmark uses the finite Kuramoto phase-oscillator model (Kuramoto, 1984) on one fixed Watts–Strogatz graph (Watts & Strogatz, 1998). Let N_{osc} be the oscillator count, k_{WS} the initial ring-lattice degree, p_{rew} the graph rewiring probability, and K_{kur} the coupling constant. The phase of node 64 is fixed to zero to remove the global phase gauge. For each remaining oscillator, θ_i is a sampled phase and ω_i the implied natural frequency. The adjacency matrix is denoted by A. The standardized learning problem uses the first 63 frequencies as physical parameters and the first 63 phases as physical states:

$$N_{\mathrm{osc}} = 64, \qquad k_{\mathrm{WS}} = 6, \qquad p_{\mathrm{rew}} = 0.2, \qquad K_{\mathrm{kur}} = 1, \tag{87}$$

$$\theta_{64} = 0, \qquad \theta_i \sim \mathrm{Unif}[-0.30, 0.30], \qquad i = 1, \ldots, 63, \tag{88}$$

$$\omega_i = K_{\mathrm{kur}} \sum_{j=1}^{64} A_{ij} \sin(\theta_i - \theta_j), \qquad \omega_{64} = -\sum_{i=1}^{63} \omega_i, \tag{89}$$

$$q = (\omega_1, \ldots, \omega_{63}), \qquad z = (\theta_1, \ldots, \theta_{63}), \qquad d = m = 63. \tag{90}$$

Sampling phases first and evaluating the equilibrium equations yields consistent natural frequencies by construction. This inverse equilibrium viewpoint is standard for finite Kuramoto systems and avoids accepting arbitrary non-equilibrium labels (Strogatz, 2000).

P3: nonlinear heat conduction. The PDE benchmark is posed on the unit-square domain Ω. The temperature field is denoted by T, and homogeneous Dirichlet data are imposed on the boundary. First-order finite elements discretize the interior into $N_{\mathrm{int}} = 529$ degrees of freedom, following the standard finite-element construction for elliptic problems (Brenner & Scott, 2008). The parameter vector contains 16 coefficients q_{ab} indexed by $a, b \in \{1, 2, 3, 4\}$, and $\xi = (\xi_1, \xi_2)$ denotes the spatial coordinate. These coefficients define the source field $f(\xi; q)$. The continuous benchmark equation and its parameterization are given below. To keep the nonlinear diffusion law explicit without overloading a single display line, let $\mathbf{j}_T$ denote the temperature-dependent conductive flux. It is defined before the

Table 8: Final hyperparameters for the independent NRGD allocation-only interface study. R is retained response rank; $h_\rho^\star$ and $h_\pi^\star$ are the finite scales for intensity and allocation; K_H and K_S are Hutchinson and geometry-direction counts. The executed full-Jacobian estimator uses $K_H = 0$. T_β and the separate mixing coefficient α are not part of this allocation-only experiment.

System	R	$h_\rho^\star$	$h_\pi^\star$	K_H	K_S	λ_ρ	λ_π	T_β	α	lr	#Params.
Synthetic $50 \to 50$	5	0.02	0.02	0	12	1	10	N/A	N/A	10^{-3}	27,399
IEEE-118 rich	6	0.02	0.02	0	12	1	10	N/A	N/A	10^{-3}	26,376
Coupled Duffing	12	0.02	0.02	0	16	1	10	N/A	N/A	10^{-3}	28,430
2D Allen–Cahn	14	0.02	0.02	0	12	1	10	N/A	N/A	10^{-3}	36,112
2D shallow-water	16	0.02	0.02	0	16	1	10	N/A	N/A	10^{-3}	37,650

The router uses $d_{\text{feat}} = d + 5R + 2$ and $\#\text{Params} = 128(d_{\text{feat}} + 1) + 128(128 + 1) + (128 + 1) + (128 + 1)(R + 1)$. Geometry-center budgets / benchmark seeds are 1024/2025 (Synthetic), 1024/4321 (IEEE-118 rich), 1024/6101 (Duffing), 1536/6102 (Allen–Cahn), and 1280/6103 (shallow-water).

steady balance equation is imposed:

$$\Omega = [0,1]^2, \qquad T|_{\partial\Omega} = 0, \qquad N_{\text{int}} = 23^2 = 529, \tag{91}$$

$$q = (q_{ab})_{a,b=1}^4, \qquad q_{ab} \sim \text{Unif}[-1,1], \qquad \xi = (\xi_1, \xi_2) \in \Omega, \tag{92}$$

$$f(\xi; q) = 10 \left[1 + 0.05 \sum_{a=1}^{4} \sum_{b=1}^{4} q_{ab} \sin(a\pi\xi_1) \sin(b\pi\xi_2) \right], \tag{93}$$

$$\mathbf{j}_T = (1 + T^2)\nabla T, \tag{94}$$

$$-\nabla \cdot \mathbf{j}_T + 0.1T^3 = f(\xi; q). \tag{95}$$

All claims concern this fixed finite-dimensional discretization. The experiment is not a mesh-convergence study and therefore makes no claim about discretization-independent continuum accuracy.

Auxiliary module-level systems. The independent NRGD and GANR studies additionally use the Synthetic and Controlled systems defined in Appendix E.2, together with Duffing, Allen–Cahn, and shallow-water. These experiments serve module-level geometry and representation diagnostics rather than the primary end-to-end physical claims. Appendix B specifies the NRGD dimensions, geometry budgets, and input settings, while Appendix C specifies the GANR data splits, probe sets, and matched model interfaces. The primary end-to-end claims of the paper are restricted to P1–P3 above.

F Experiment-specific Protocols

F.1 NRGD interface-fidelity protocol

The independent NRGD experiment uses five implicit systems: Synthetic, IEEE-118, Duffing, Allen–Cahn, and shallow-water. An input-only descriptor predictor is evaluated on a held-out split with three seeds. Its input setting is compared with raw-input, random-projector, intensity-only, and training-mean controls. A directional-response sketch with additional information is included only as an interface diagnostic and is not treated as an inference-time baseline. The endpoints are complete allocation MSE, total-response MSE, and fraction NMSE; lower values are better for all three, with the two MSE columns scaled by 10^{-3} in Table 1. The target is an NRGD system-side descriptor derived from finite-scale response geometry, not the final physical state prediction.

F.2 GANR factorial response-representation protocol

The independent GANR factorial uses Controlled, Duffing, IEEE-118, Allen–Cahn, and shallow-water with seeds $\{11, 29, 47\}$. The compared models are ReLU, GELU, SiLU, SwiGLU, ResMLP, Fourier, and the auxiliary response-aware GANR formulation. Each model is evaluated under two auxiliary conditions: No physics, in which the equation-derived linearization is absent, and Shared

physics, in which the same fixed linearization is supplied to every model. The primary endpoint is standardized NRMSE, reported in Table 2; lower is better. For each model the upper table row is the no-physics condition and the lower row is the shared-physics condition. The comparison is interpreted within condition, so differences caused by the auxiliary physics factor are not conflated with the effect of response-aware nonlinear computation. Appendix C gives the data splits, probe sets, response-stratified diagnostics, and strict plug-in experiment. As stated there, this auxiliary GANR formulation tests the representation premise and is not identical to the final dual-subspace, layer-wise adapter used by GA-NAR.

F.3 GA-NAR END-TO-END PROTOCOL

The end-to-end experiment uses P1 IEEE-118 AC power flow, P2 Kuramoto-64, and P3 nonlinear heat FEM. Six activation families are paired with activation-matched baselines over three seeds, giving $6 \times 3 \times 2 \times 3 = 108$ model trainings. Each pair shares the system split, seed, minibatch order, and 400-epoch budget. Architecture and capacity differences are stated once in Appendix G. The reported three-seed means and standard deviations are descriptive; primary endpoints are standardized test MSE and independently reevaluated physical-equation residuals (Table 3).

G BACKBONE PAIRING AND CAPACITY CONTEXT

The primary end-to-end comparison is organized as matched activation–system pairs. Six activation families are evaluated on three physical systems with three seeds per cell, yielding 108 trainings. Within each pair, the data split, seed, minibatch ordering, and epoch budget are identical. Residual pathways follow the standard residual-network construction (He et al., 2016); GELU, SiLU/Swish, Mish, and SwiGLU follow their standard definitions (Hendrycks & Gimpel, 2016; Ramachandran et al., 2018; Misra, 2020; Shazeer, 2020).

Each activation-matched baseline contains four blocks with hidden width 128 and feed-forward width 256. The GA-NAR-augmented backbone contains four geometry-modulated blocks with hidden width 192, feed-forward width 384, geometry rank 96, and the normalized low-rank modulation path defined above. In the paired evaluation, the GANR-modulated backbone serves as a trainable residual-correction branch whose output is added to a fixed activation-matched baseline predictor. The correction is initialized by ridge regression on training residuals, with its regularization coefficient selected using validation data only.

Because the augmented backbone has additional trainable capacity, Table 3 is interpreted as an end-to-end comparison of the complete GA-NAR method rather than as parameter-matched architectural superiority. Appendix B isolates the validity of the system-side response geometry, and Appendix C tests response-aware nonlinear representation under shared interfaces; neither is treated as a parameter-matched ablation of the complete end-to-end method. With three seeds per cell, means, standard deviations, and paired seed-level wins are descriptive rather than formal significance tests, and transfer claims are limited to the tested activation families and system classes.

H CONTROLLED ATTRIBUTION OF GA-NAR GAINS

The primary end-to-end table compares the complete GA-NAR augmentation with a smaller activation-matched baseline, so it cannot by itself distinguish gains from extra correction capacity, access to NRGD information, and the structured GANR mechanism. We therefore run a dedicated four-way controlled ablation on seven representative system–activation settings using the same three seeds $\{0, 1, 2\}$ and the same 400-epoch training budget. The four methods are: (i) the Original Baseline; (ii) Capacity Control, which adds the same correction-branch budget as GA-NAR but uses only x; (iii) Generic Geometry Control, which uses the same correction capacity and the same fixed NRGD information through

$$c(x) = [x,\, P_R x,\, (I - P_R)x,\, \rho_G^\star(x),\, \boldsymbol{\pi}_G^\star(x)] \tag{96}$$

inside a generic residual correction network, without layer-wise modulation, depth allocation, or scale matching; and (iv) Full GA-NAR, which uses the structured dual-subspace and layer-wise geometry modulation. Thus the comparison separates extra capacity, access to geometry information, and the

Table 9: Four-way controlled attribution: standardized test MSE. Each value row gives mean ± sample SD over seeds $\{0, 1, 2\}$; the row immediately below gives paired improvement versus the Original Baseline for the three controlled variants. The Original Baseline is shown once and vertically spans the value/improvement pair. System labels state the applied scale. Lower MSE and larger positive improvement are better.

System	Activation	Original Baseline	Capacity Control	Generic Geometry Control	Full GA-NAR
IEEE-118 ($\times 10^{-6}$)	ReLU	2.587 ± 0.210	2.453 ± 0.178 $5.15 \pm 0.87\%$	2.548 ± 0.197 $1.50 \pm 0.45\%$	2.376 ± 0.182 $8.13 \pm 0.58\%$
	GELU	3.990 ± 0.072	3.496 ± 0.074 $12.38 \pm 0.34\%$	3.659 ± 0.085 $8.32 \pm 0.74\%$	3.529 ± 0.070 $11.56 \pm 0.17\%$
	SwiGLU	1.825 ± 0.084	1.472 ± 0.053 $19.29 \pm 1.80\%$	1.552 ± 0.040 $14.90 \pm 2.48\%$	1.443 ± 0.054 $20.90 \pm 1.70\%$
Kuramoto-64 ($\times 10^{-3}$)	ReLU	18.677 ± 0.250	1.002 ± 0.115 $94.64 \pm 0.60\%$	0.953 ± 0.095 $94.90 \pm 0.50\%$	1.055 ± 0.128 $94.36 \pm 0.64\%$
	GELU	0.325 ± 0.041	0.321 ± 0.040 $1.12 \pm 0.66\%$	0.321 ± 0.040 $1.11 \pm 0.67\%$	0.319 ± 0.040 $1.83 \pm 0.56\%$
Heat FEM ($\times 10^{-6}$)	ReLU	2.393 ± 0.071	1.288 ± 0.103 $46.22 \pm 2.68\%$	1.291 ± 0.103 $46.11 \pm 2.67\%$	1.283 ± 0.104 $46.46 \pm 2.74\%$
	Softplus	4.461 ± 0.063	1.946 ± 0.026 $56.37 \pm 0.69\%$	1.946 ± 0.028 $56.38 \pm 0.70\%$	0.461 ± 0.045 $89.67 \pm 1.14\%$

Table 10: Four-way controlled attribution: independently recomputed physical-equation residual mean. Each value row gives mean ± sample SD over seeds $\{0, 1, 2\}$; the row immediately below gives paired improvement versus the Original Baseline for the three controlled variants. The Original Baseline is shown once and vertically spans the value/improvement pair. System labels state the applied scale. Lower residual and larger positive improvement are better.

System	Activation	Original Baseline	Capacity Control	Generic Geometry Control	Full GA-NAR
IEEE-118 ($\times 10^{-3}$)	ReLU	8.031 ± 0.551	7.144 ± 0.282 $10.91 \pm 2.96\%$	7.387 ± 0.360 $7.90 \pm 3.65\%$	6.990 ± 0.312 $12.87 \pm 2.15\%$
	GELU	10.615 ± 0.128	9.207 ± 0.112 $13.25 \pm 1.86\%$	9.408 ± 0.156 $11.37 \pm 1.72\%$	9.324 ± 0.157 $12.16 \pm 1.11\%$
	SwiGLU	8.546 ± 0.082	6.275 ± 0.305 $26.59 \pm 3.00\%$	6.575 ± 0.442 $23.08 \pm 4.65\%$	6.271 ± 0.204 $26.62 \pm 1.97\%$
Kuramoto-64 ($\times 10^{-3}$)	ReLU	39.176 ± 2.220	26.391 ± 1.900 $32.49 \pm 6.27\%$	25.201 ± 1.620 $35.53 \pm 5.78\%$	27.287 ± 1.970 $30.24 \pm 5.82\%$
	GELU	12.021 ± 0.083	11.977 ± 0.059 $0.36 \pm 0.21\%$	11.981 ± 0.055 $0.33 \pm 0.24\%$	11.950 ± 0.063 $0.59 \pm 0.18\%$
Heat FEM ($\times 10^{-5}$)	ReLU	4.011 ± 0.114	3.333 ± 0.123 $16.92 \pm 0.76\%$	3.337 ± 0.124 $16.82 \pm 0.80\%$	3.321 ± 0.123 $17.23 \pm 0.81\%$
	Softplus	7.196 ± 0.047	4.648 ± 0.030 $35.41 \pm 0.50\%$	4.647 ± 0.033 $35.43 \pm 0.52\%$	2.335 ± 0.086 $67.55 \pm 1.35\%$

GA-NAR-specific realization mechanism. All physical residuals are recomputed independently from the governing equations and are not training losses.

For a metric E and seed i, paired improvement relative to the Original Baseline is

$$\Delta_i(E) = \frac{E_{\mathrm{base,i}} - E_{\mathrm{method,i}}}{E_{\mathrm{base,i}}}. \tag{97}$$

Tables 9 and 10 report the mean ± sample SD over seeds; the second line in each non-baseline cell is the paired improvement $100\ \mathrm{mean}(\Delta_i) \pm 100\ \mathrm{SD}(\Delta_i)$.

The controls show that extra capacity is a genuine contributor: Capacity Control improves both MSE and physical residual over the Original Baseline in all seven settings. Generic access to the same NRGD information also improves the baseline in all seven settings. Crucially, however, neither factor reproduces the full pattern of the structured method. Full GA-NAR gives larger paired MSE improvements than Capacity Control in five of seven settings and than Generic Geometry Control in

six of seven; the same counts hold for physical-residual improvement. The distinction is especially clear for nonlinear heat FEM with Softplus, where Full GA-NAR reaches 89.67% paired MSE improvement and 67.55% paired residual improvement, compared with approximately 56.4% and 35.4% for the two matched controls.

The result therefore rules out extra correction capacity or mere access to NRGD descriptors as a sole explanation for the observed gains. It does not support a stronger claim that structured modulation must dominate in every setting: Capacity Control is slightly better for IEEE-118 with GELU, and Generic Geometry Control is slightly better for Kuramoto-64 with ReLU. The controlled evidence instead supports the narrower causal statement needed by the paper: the GA-NAR-specific dual-subspace, layer-wise realization contributes beyond capacity and geometry access in the majority of representative settings, while the magnitude of that contribution is activation- and system-dependent.

I Reported Metrics and State-Dependence Diagnostics

Prediction and physical consistency are evaluated on the independent supervised test set $\mathcal{D}_{\text{test}}$. Let m be the standardized output dimension and let $\widehat{G}$ denote the evaluated predictor, whether baseline or GA-NAR. The standardized mean-squared error averages coordinate-wise prediction error, while the sample-wise physical residual $R_{\text{phys}}(x)$ uses the scaled governing residual $\bar{F}$ defined in Appendix A. Its test-set average is $\overline{R}_{\text{phys}}$:

$$\text{MSE}_{\text{std}} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{x \in \mathcal{D}_{\text{test}}} \frac{1}{m} \|\widehat{G}(x) - G(x)\|_2^2, \tag{98}$$

$$R_{\text{phys}}(x) = \frac{\|\bar{F}(\widehat{G}(x), x)\|_2}{\sqrt{m}}, \tag{99}$$

$$\overline{R}_{\text{phys}} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{x \in \mathcal{D}_{\text{test}}} R_{\text{phys}}(x). \tag{100}$$

A non-finite prediction is recorded as a numerical failure and receives infinite sample-level error. The residual is a consistency diagnostic rather than a certificate of global branch membership or satisfaction of inequality constraints. Metrics specific to the independent NRGD and GANR studies are defined in their experiment protocols and table captions.

State-dependence diagnostics. The independent A/B geometry split tests whether local NRGD descriptors vary across operating states beyond stochastic-probe disagreement. For intensity, let $\rho_G^{\star,A}$ and $\rho_G^{\star,B}$ be the two split estimates and let m_ρ be their mean. The quantity $\sigma^2_{\rho,\text{diff}}$ measures split disagreement, while $\sigma^2_{\rho,\text{sig}}$ estimates between-state signal after subtracting half of that disagreement variance; S_ρ is their signal-to-disagreement ratio. For allocation, let $p_G^A = \pi_R^{\star,A}$ and $p_G^B = \pi_R^{\star,B}$ be the dominant-share estimates, with analogous quantities m_π, $\sigma^2_{\pi,\text{diff}}$, $\sigma^2_{\pi,\text{sig}}$, and S_π. Let ϵ_{64} denote FP64 machine epsilon. We use

$$m_\rho(x) = \frac{\rho_G^{\star,A}(x) + \rho_G^{\star,B}(x)}{2}, \qquad \sigma^2_{\rho,\text{diff}} = \frac{1}{2}\mathbb{E}_x[(\rho_G^{\star,A} - \rho_G^{\star,B})^2], \tag{101}$$

$$\sigma^2_{\rho,\text{sig}} = \max\left\{ \text{Var}_x(m_\rho) - \frac{\sigma^2_{\rho,\text{diff}}}{2}, 0 \right\}, \qquad S_\rho = \frac{\sigma_{\rho,\text{sig}}}{\sqrt{\sigma^2_{\rho,\text{diff}}/2 + 100\epsilon_{64}}}, \tag{102}$$

$$m_\pi(x) = \frac{p_G^A(x) + p_G^B(x)}{2}, \qquad \sigma^2_{\pi,\text{diff}} = \frac{1}{2}\mathbb{E}_x[(p_G^A - p_G^B)^2], \tag{103}$$

$$\sigma^2_{\pi,\text{sig}} = \max\left\{ \text{Var}_x(m_\pi) - \frac{\sigma^2_{\pi,\text{diff}}}{2}, 0 \right\}, \qquad S_\pi = \frac{\sigma_{\pi,\text{sig}}}{\sqrt{\sigma^2_{\pi,\text{diff}}/2 + 100\epsilon_{64}}}. \tag{104}$$

The geometry-resolution gate in Appendix J requires these state-dependence statistics to exceed their prespecified thresholds.

J PRESPECIFIED EXPERIMENTAL PROTOCOL

Regularity, admissibility, and continuation. All physical systems are solved in FP64 with damped Newton iterations and Armijo backtracking. Armijo's sufficient-decrease rule is used only as a numerical globalization mechanism (Armijo, 1966). Let c_1 be the Armijo constant, β_{ls} the multiplicative line-search contraction factor, $N_{\mathrm{ls}}^{\max}$ the maximum number of line-search reductions, and $N_{\mathrm{Newton}}^{\max}$ the maximum number of Newton updates. Let $\|\bar{F}\|_{\mathrm{tol}}$ denote the scaled-residual tolerance, let $\mathrm{rcond}_1(\bar{F}_y)$ denote the reciprocal 1-norm condition estimate of the state Jacobian, and let N_{cont} be the number of continuation substeps used for a finite-scale endpoint. Their fixed values are

$$c_1 = 10^{-4}, \qquad \beta_{\mathrm{ls}} = 0.5, \qquad N_{\mathrm{ls}}^{\max} = 12, \qquad N_{\mathrm{Newton}}^{\max} = 20, \tag{105}$$

$$\|\bar{F}\|_{\mathrm{tol}} = 10^{-12}, \qquad \mathrm{rcond}_1(\bar{F}_y) \geq 10^{-8}, \qquad N_{\mathrm{cont}} = 4. \tag{106}$$

Geometry centers are never clipped or reflected at parameter boundaries. A center–direction–scale tuple is valid only when every continuation state is finite, admissible, regular, inside the prescribed parameter domain, and connected to the center through the specified continuation path. The same rule is used in derivative certification and geometry construction.

Derivative certification. Every benchmark is certified at three frozen states before NRGD is executed. Let $\mathrm{Rel}(a, b)$ denote the symmetric relative discrepancy between two vectors. For residual derivatives, $v_y \in \mathbb{R}^m$ and $v_x \in \mathbb{R}^d$ are frozen output- and input-direction probes. Their analytic actions are a_y and a_x, while $d_y^{(\delta)}$ and $d_x^{(\delta)}$ are centered finite-difference references at step size δ. The residual finite-difference grid is denoted by $\delta_{\mathrm{residual}}$. The definitions and acceptance conditions are

$$\mathrm{Rel}(a, b) = \frac{\|a - b\|_2}{\|a\|_2 + \|b\|_2 + 10^{-12}}, \tag{107}$$

$$a_y = \bar{F}_y(y, x)v_y, \qquad d_y^{(\delta)} = \frac{\bar{F}(y + \delta v_y, x) - \bar{F}(y - \delta v_y, x)}{2\delta}, \tag{108}$$

$$a_x = \bar{F}_x(y, x)v_x, \qquad d_x^{(\delta)} = \frac{\bar{F}(y, x + \delta v_x) - \bar{F}(y, x - \delta v_x)}{2\delta}, \tag{109}$$

$$\delta_{\mathrm{residual}} \in \{10^{-4}, 10^{-5}, 10^{-6}\}, \tag{110}$$

$$\min_{\delta} \mathrm{Rel}(a_y, d_y^{(\delta)}) \leq 10^{-6}, \qquad \min_{\delta} \mathrm{Rel}(a_x, d_x^{(\delta)}) \leq 10^{-6}. \tag{111}$$

The solution-map JVP is certified separately. Let j_{IFT} denote the implicit JVP and $j_{\mathrm{solve}}^{(\delta)}$ the centered finite-difference reference obtained from two physical solutions. The solution-level step grid is $\delta_{\mathrm{solution}}$:

$$j_{\mathrm{IFT}} = J_G(x)v_x, \qquad j_{\mathrm{solve}}^{(\delta)} = \frac{G(x + \delta v_x) - G(x - \delta v_x)}{2\delta}, \tag{112}$$

$$\delta_{\mathrm{solution}} \in \{10^{-3}, 10^{-4}, 10^{-5}\}, \tag{113}$$

$$\min_{\delta} \mathrm{Rel}(j_{\mathrm{IFT}}, j_{\mathrm{solve}}^{(\delta)}) \leq 10^{-5}. \tag{114}$$

At least one adjacent pair of step sizes must additionally satisfy the stability tolerance of 10^{-4}. Every finite-difference endpoint must satisfy the same admissibility checks as a geometry probe. Failed certification marks downstream geometry unresolved rather than activating an alternative derivative path. The residual-step grid and its two acceptance inequalities are stated separately for clarity.

Finite-scale selection. The candidate finite-scale set is denoted by $\mathcal{H}$. The training-only geometry set is divided into disjoint halves indexed by $c \in \{A, B\}$. For adjacent candidate indices i and $i + 1$, $\mathcal{D}_{i,i+1}^{c}$ is the subset of centers valid at both scales, and $f_{i,i+1}^{c}$ is its coverage fraction relative to the geometry half $\mathcal{D}_{\mathrm{geo}}^{c}$. The empirical response moment at scale h is $\widehat{S}_{c,i}^{h}$. For center x, $s_{x,1}$ denotes its first frozen unit-sphere response direction, and v_ℓ denotes Hutchinson probe ℓ. The quantity $\widehat{J}_{\mathrm{rms},c,i}^{2}$ measures the mean squared Jacobian-action magnitude on the same centers, and $\eta_{\mathrm{NL},c,i}^{h}$ is the dimensionless nonlinear-response ratio. The candidate scales, geometry-set sizes, coverage

requirement, and response-signal requirement are

$$\mathcal{H} = \{0.02, 0.01, 0.005, 0.0025\}, \qquad N_{\text{geo}}^{\text{P1}} = 1024, \quad N_{\text{geo}}^{\text{P2}} = N_{\text{geo}}^{\text{P3}} = 512, \tag{115}$$

$$f_{i,i+1}^{c} = \frac{|\mathcal{D}_{i,i+1}^{c}|}{|\mathcal{D}_{\text{geo}}^{c}|} \geq 0.95, \qquad c \in \{A, B\}, \tag{116}$$

$$\widehat{S}_{c,i}^{h} = \frac{1}{|\mathcal{D}_{i,i+1}^{c}|} \sum_{x \in \mathcal{D}_{i,i+1}^{c}} \widehat{r}_{G}^{h}(x, s_{x,1}) s_{x,1} s_{x,1}^{\top}, \tag{117}$$

$$\widehat{J}_{\text{rms},c,i}^{2} = \frac{1}{|\mathcal{D}_{i,i+1}^{c}|} \sum_{x \in \mathcal{D}_{i,i+1}^{c}} \frac{1}{K_H} \sum_{\ell=1}^{K_H} \|J_G(x) v_\ell\|_2^2, \tag{118}$$

$$\eta_{\text{NL},c,i}^{h} = \frac{\text{tr}(\widehat{S}_{c,i}^{h})}{\widehat{J}_{\text{rms},c,i}^{2} + \epsilon_{64}} \geq 10^{-8}. \tag{119}$$

The two NRGD target families may use different selected scales because they test different aspects of finite-scale stability. For intensity, let $a_{c,i}^{h}$ be the total response magnitude and let $E_{\text{scale,int},c}^{i,i+1}$ be its adjacent-scale relative discrepancy. For allocation, $\Gamma_{c,i}^{h}$ is the trace-normalized response moment, $C_{c,i}^{h}$ its anisotropic component after subtracting the isotropic identity term, and $E_{\text{scale,alloc},c}^{i,i+1}$ the adjacent-scale shape discrepancy. The selection quantities are

$$a_{c,i}^{h} = \text{tr}(\widehat{S}_{c,i}^{h}), \qquad \Gamma_{c,i}^{h} = \frac{\widehat{S}_{c,i}^{h}}{\text{tr}(\widehat{S}_{c,i}^{h})}, \qquad C_{c,i}^{h} = \Gamma_{c,i}^{h} - \frac{I_d}{d}, \tag{120}$$

$$E_{\text{scale,int},c}^{i,i+1} = \frac{2|a_{c,i}^{h_i} - a_{c,i}^{h_{i+1}}|}{a_{c,i}^{h_i} + a_{c,i}^{h_{i+1}}}, \tag{121}$$

$$\|C_{c,i}^{h_i}\|_F + \|C_{c,i}^{h_{i+1}}\|_F > 100\epsilon_{64}, \tag{122}$$

$$E_{\text{scale,alloc},c}^{i,i+1} = \frac{2\|C_{c,i}^{h_i} - C_{c,i}^{h_{i+1}}\|_F}{\|C_{c,i}^{h_i}\|_F + \|C_{c,i}^{h_{i+1}}\|_F}, \tag{123}$$

$$E_{\text{scale,int},c}^{i,i+1} \leq 0.25, \qquad E_{\text{scale,alloc},c}^{i,i+1} \leq 0.25. \tag{124}$$

Split A selects the first admissible adjacent-scale pair from large to small, and split B must independently confirm that pair. The larger scale in the first confirmed pair is selected for the corresponding target family. No scale is reselected after neural results are observed.

Geometry-resolution gate. The Geometry Resolution Gate separates numerical prerequisites from substantive repeatability and state-dependence requirements. For the global subspace, E_{proj} is the relative discrepancy introduced by projecting $\widehat{K}_G$ to $\widehat{M}_G$; g_R is the normalized eigengap at the selected rank; Q_{sub} is the overlap between the independent A/B projectors P_R^A and P_R^B; E_{sym}, E_{idem}, and E_{tr} are symmetry, idempotence, and trace-consistency diagnostics; and f_{moment} is the valid-center fraction used by the response moment. The global gate is

$$E_{\text{proj}} = \frac{\|\widehat{M}_G - \widehat{K}_G\|_F}{\|\widehat{K}_G\|_F + 100\epsilon_{64}} \leq 0.25, \tag{125}$$

$$1 \leq R < d, \qquad g_R = \frac{\gamma_R - \gamma_{R+1}}{\gamma_1 + 100\epsilon_{64}} \geq 10^{-8}, \tag{126}$$

$$Q_{\text{sub}} = \frac{1}{R} \text{tr}(P_R^A P_R^B) \geq 0.90, \tag{127}$$

$$E_{\text{sym}}, E_{\text{idem}}, E_{\text{tr}} \leq 10^{-10}, \qquad f_{\text{moment}} \geq 0.95. \tag{128}$$

Local intensity is evaluated only after the global prerequisites pass. Let $\mathcal{D}_\rho$ be the centers with valid intensity targets, f_ρ their coverage fraction, and Q_{int} the mean A/B agreement of the bounded

intensity target. The state-dependence statistic S_ρ was defined above. The local intensity gate is

$$f_\rho = \frac{|\mathcal{D}_\rho|}{N_{\mathrm{geo}}}, \qquad Q_{\mathrm{int}} = 1 - \mathbb{E}_{x\in\mathcal{D}_\rho}|\rho_G^{\star,A}(x) - \rho_G^{\star,B}(x)|, \tag{129}$$

$$f_\rho \geq 0.95, \qquad Q_{\mathrm{int}} \geq 0.90, \qquad S_\rho \geq 2. \tag{130}$$

For allocation, let $\epsilon_{\mathrm{alloc}}$ be the minimum accepted total subspace-response level, let $\mathcal{D}_\pi$ be the valid allocation-target set, and let $\mathcal{D}_\pi^{AB}$ be the subset valid in both A/B constructions. Their coverage fractions are f_{alloc} and f_{alloc}^{AB}. The A/B allocation agreement is Q_{alloc}, and S_π is the state-dependence statistic defined above. The minimum signal ratios for intensity and allocation are denoted by $\eta_\rho^{\min}$ and $\eta_{\mathrm{alloc}}^{\min}$. The allocation and final scale checks are

$$\epsilon_{\mathrm{alloc}} = 10^{-8} d\ \mathrm{tr}(\widehat{S}_G), \qquad \tau_{G,R} + \tau_{G,\perp} > \epsilon_{\mathrm{alloc}}, \tag{131}$$

$$f_{\mathrm{alloc}} = \frac{|\mathcal{D}_\pi|}{N_{\mathrm{geo}}}, \qquad f_{\mathrm{alloc}}^{AB} = \frac{|\mathcal{D}_\pi^{AB}|}{N_{\mathrm{geo}}}, \tag{132}$$

$$Q_{\mathrm{alloc}} = 1 - \mathbb{E}_{x\in\mathcal{D}_\pi^{AB}} H^2(\boldsymbol{\pi}_G^{\star,A}(x), \boldsymbol{\pi}_G^{\star,B}(x)), \tag{133}$$

$$f_{\mathrm{alloc}} \geq 0.95, \qquad f_{\mathrm{alloc}}^{AB} \geq 0.95, \qquad Q_{\mathrm{alloc}} \geq 0.90, \qquad S_\pi \geq 2, \tag{134}$$

$$\eta_\rho^{\min} \geq 10^{-8}, \qquad \eta_{\mathrm{alloc}}^{\min} \geq 10^{-8}, \tag{135}$$

$$E_{\mathrm{scale,int}}^{\max} \leq 0.25, \qquad E_{\mathrm{scale,alloc}}^{\max} \leq 0.25. \tag{136}$$

All continuous diagnostics are computed before the final binary gate. Failure of any prerequisite marks the associated geometry quantity unresolved. A failed criterion leaves the corresponding target unresolved, and thresholds are not changed after observing neural results.

Data, neural optimization, and held-out geometry. The supervised split and geometry-probe split are fixed before model training. For system P1, N_{train}, N_{val}, and N_{test} denote supervised training, validation, and test counts. The same symbols are used for P2 and P3 with system-specific values. The quantities N_{geo} and N_{probe} denote training-only geometry centers and disjoint held-out geometry-probe centers. Their fixed sizes are

$$(N_{\mathrm{train}}, N_{\mathrm{val}}, N_{\mathrm{test}})_{\mathrm{P1}} = (20000, 2000, 2000), \tag{137}$$

$$(N_{\mathrm{train}}, N_{\mathrm{val}}, N_{\mathrm{test}})_{\mathrm{P2}} = (N_{\mathrm{train}}, N_{\mathrm{val}}, N_{\mathrm{test}})_{\mathrm{P3}} = (10000, 1000, 1000), \tag{138}$$

$$(N_{\mathrm{geo}}, N_{\mathrm{probe}})_{\mathrm{P1}} = (1024, 512), \qquad (N_{\mathrm{geo}}, N_{\mathrm{probe}})_{\mathrm{P2}} = (N_{\mathrm{geo}}, N_{\mathrm{probe}})_{\mathrm{P3}} = (512, 256). \tag{139}$$

Probe centers are excluded from neural parameter fitting, early stopping, geometry scale/rank selection, and the independent supervised test set. Optimization uses AdamW (Loshchilov & Hutter, 2019). Let β_1 and β_2 be its moment coefficients, ϵ_{opt} its numerical stabilizer, wd the weight-decay coefficient, B the minibatch size, $N_{\mathrm{epoch}}^{\max}$ the maximum epoch count, $\|g\|_{2,\max}$ the gradient-norm clipping threshold, $\mathrm{patience}$ the early-stopping patience, and $\mathrm{lr}_{\min}$ the minimum learning rate under cosine decay. All reported paired runs use the fixed seed set $\mathcal{S} = \{0, 1, 2\}$:

$$\beta_1 = 0.9, \qquad \beta_2 = 0.999, \qquad \epsilon_{\mathrm{opt}} = 10^{-8}, \qquad \mathrm{wd} = 10^{-5}, \tag{140}$$

$$B = 256, \qquad N_{\mathrm{epoch}}^{\max} = 400, \qquad \|g\|_{2,\max} = 1, \qquad \mathrm{patience} = 50, \tag{141}$$

$$\mathrm{lr}_{\min} = 10^{-5}, \qquad \mathcal{S} = \{0, 1, 2\}. \tag{142}$$

The maximum learning rate and other hyperparameters are selected from validation data only; the independent test set is not used for selection. Automatic mixed precision and TF32 are disabled, and the validation-selected model is evaluated in FP64.

Randomness and run budget. All stochastic components are generated deterministically from a master seed using disjoint random-number streams. Let s_{master} denote the master seed and N_{paired} the number of trainings in the paired activation–system matrix:

$$s_{\mathrm{master}} = 20260924, \qquad N_{\mathrm{paired}} = 6 \times 3 \times 2 \times 3 = 108. \tag{143}$$

Dataset splits, training statistics, geometry sets, probe banks, selected scales, response rank, projector, initialization, and minibatch orders are fixed before test evaluation. No replacement seed or best-seed selection is permitted. The three-seed results are summarized descriptively without a formal significance claim.

The physical solver, Jacobian actions, and NRGD target construction are used only to build training-time geometry descriptors and independent evaluation diagnostics. Inference is fully neural and does not query the physical Jacobian, JVPs, or NRGD target values.