WHEN BOUNDARY STRUCTURE HELPS: A CONTROLLED STUDY OF NEURAL PDE OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Boundary conditions provide inputs to neural operators for partial differential equations (PDEs), but can score structure improve learning when information and access are matched? We study the Boundary Heterophily Operator (BHO), a quadratic residual on boundary-key logits. On the Boundary-Embedded Neural Operators (BENO) benchmark with 100 examples used for gradient training, BHO reduces held-out relative-L_2 error by 13.87% on the source shape and 8.27% under cross-shape transfer. The low-data advantage persists under two matched training budgets. The full quadratic also outperforms a cross-only score with the same boundary conditioner in BENO and a fixed-source Dirichlet circle problem. A secondary metric computed from the same circle predictions favors BHO over the host, a boundary-group bias and the cross-only score in predicting boundary-induced solution changes. Native mixed-boundary comparisons do not establish BHO's superiority over the host or a scalar message gate, and the tested elasticity comparison favors these controls. In a separate study with constant conditioning and physical-interior queries, removing within-head sign mixing increases error by 23.19% with a shared query/key residual projection but has no detected cost with separate projections. Thus boundary-score parameterization can improve learning within a matched host and route, while sign-restriction costs depend on the tested residual parameterization.

1 INTRODUCTION

A boundary-value problem seeks a function satisfying a partial differential equation (PDE) inside a domain and prescribed conditions on its boundary. For Poisson's equation, these may specify the solution value (Dirichlet), its normal derivative (Neumann), or their linear combination (Robin) (Mousavi et al., 2026). When boundary conditions vary across examples, a learned operator must use them as sample-dependent inputs. Neural operators learn maps between input and solution functions (Kovachki et al., 2023); here the inputs comprise the source function, boundary data and domain geometry.

Existing approaches differ in how they represent and access boundary data. Boundary-Embedded Neural Operators (BENO) form a global boundary embedding (Wang et al., 2024); the General Neural Operator Transformer (GNOT) introduces input functions through cross-attention (Hao et al., 2023); and the Physics-Informed Transformer Neural Operator (PINTO) conditions operator learning on initial and boundary data (Boya & Subramani, 2025). Learned function extensions map boundary data into domain-level features (Mousavi et al., 2026), including for geometry-aware operator transformers (Wen et al., 2025). Other approaches use operator constraints, embedded physics or boundary-adapted spectral representations (Saad et al., 2023; Horie & Mitsume, 2022; Liu et al., 2023). A further design choice concerns how an attention score uses information already supplied through a fixed access path.

Attention kernels, structural biases and signed similarities offer established ways to shape interactions (Shaw et al., 2018; Ying et al., 2021; Tsai et al., 2019; Kim et al., 2019). The Boundary Heterophily Operator (BHO) augments boundary-key logits with a boundary-conditioned quadratic residual while retaining ordinary attention and value aggregation. The feature-only host is the same neural operator without this residual and receives identical inputs. The comparisons test how boundary-score parameterization affects learning when information and access are matched.

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.

The first question (Q1) asks what this added score contributes within an information-matched host and route, and how far that evidence extends. BENO source-shape and cross-shape evaluations with 100 examples used for gradient training show a benefit beyond supplied features in the joint-attention host. This benefit persists under two matched development budgets (Section 3.1). The full quadratic also outperforms a score retaining only the query–key cross term with the same boundary conditioner in BENO and the circle population from operator learning with boundary conditions (OLBC) (Sections 3.2–3.3). For the same circle predictions, a secondary endpoint measures the accuracy of the predicted solution change when only boundary values change. Scope comparisons then examine other access designs, a released boundary-aware host at its original node resolution, and an elasticity population (Section 4). They distinguish improvement over a simpler score from benefit over the host itself.

A useful score need not require the full flexibility of its parameterization. The second question (Q2) asks whether removing within-head sign mixing has the same cost with separate query/key residual projections (directed) or one common projection (shared). This separate study uses a constant conditioner input and restricts the residual to physical-interior queries (Section 5). Removing sign mixing has a detected cost in the shared family but not the directed family. Within the shared model, fixing initial mixed signs is non-inferior to learning their assignment within a 3% margin.

2 Boundary information and controlled intervention

2.1 Supplied information and prediction task

The prediction task is to learn $(f, g, \Omega) \mapsto u$, mapping the interior source f, boundary data g, and domain Ω to the solution u. The benchmark descriptions formulate the continuum Poisson problem as $\Delta u = f$ for BENO and $-\Delta u = f$ for OLBC (Wang et al., 2024; Mousavi et al., 2026); the targets are the released numerical solutions. BENO training uses the Neumann-labeled four-corner collection, whose source, boundary data, and corner sizes vary across examples (Appendix A.5). The OLBC circle dataset, released as `poisson-circle-bc1` (bc1), fixes geometry and source while varying Dirichlet values.

In these grid-based comparisons, source values, boundary data, coordinates and geometry are supplied as token features. Boundary tokens also carry estimated normals and a boundary-type label. The feature-only host receives every feature supplied to BHO, so their comparison isolates score structure from supplied information. The prediction target is u at physical-interior grid nodes. The per-example relative error is defined as

$$E(\hat{u}, u) = \frac{\|\hat{u} - u\|_{2,\Omega}}{\|u\|_{2,\Omega} + \epsilon}, \tag{1}$$

where $\hat{u}$ is the prediction, $\|\cdot\|_{2,\Omega}$ is the discrete norm over the evaluated nodes, and $\epsilon > 0$ prevents division by zero. The reported metric averages this error over evaluation examples; the same mean is the training objective for the BENO comparisons and OLBC bc1 score controls. Native-resolution evaluation is described in Section 4.2.

2.2 A score intervention on an existing route

The host's access path determines which keys participate in attention. The primary joint-attention host (B1) attends over grid and boundary tokens jointly.

BHO modifies boundary-key logits on the host's existing path while preserving token inputs, ordinary attention and value aggregation (Figure 1). The B1 host has four pre-normalized transformer blocks, width $d = 128$, and eight heads; it processes 1,024 grid tokens and 128 boundary tokens. In layer ℓ and head h, let $q_i = Qh_i$ and $k_j = Kh_j$ be the host's query and key projections of normalized hidden states h_i, h_j, with head dimension $d_h = 16$. Layer/head indices on Q, K, W, q, k, D are suppressed; parameters are not tied across layers or heads. At boundary key j, the feature vector and diagonal conditioner are

$$e_j = [\hat{n}_{x,j}, \hat{n}_{y,j}, \mathrm{bctype}_j, g_j]^\top, \qquad D_j = \mathrm{diag}(\tanh(We_j)).$$

Here $\hat{n}_{x,j}, \hat{n}_{y,j}$ are estimated normal components, bctype_j is the boundary-type code, g_j is the scaled boundary datum, and $W \in \mathbb{R}^{d_h \times 4}$ is a learned conditioning matrix. For the set $\mathcal{B}$ of boundary-

key indices, the primary directed score is

$$s_{ij} = \frac{q_i^\top k_j}{\sqrt{d_h}} + \mathbf{1}\{j \in \mathcal{B}\}\lambda_{\ell,h}(q_i - k_j)^\top D_j(q_i - k_j). \tag{2}$$

The residual lies outside the ordinary $1/\sqrt{d_h}$ scaling and acts on every query, including boundary queries. Joint softmax normalizes over all 1,152 keys, followed by the host's ordinary value aggregation. The real-valued gates $\lambda_{\ell,h}$ start at zero, matching the feature-only host's initial predictor. The residual introduces 2,080 parameters (0.38%) beyond the 546,689-parameter host, thereby altering the learning parameterization while preserving the supplied information. In this implementation, inference and forward–loss–backward latencies are $1.56\times$ and $1.59\times$ those of the host, respectively (Appendix E.1).

For comparisons across access paths, the cross-attention host (B2) uses grid self-attention and a separate per-block boundary cross-attention route. The slice host (B3) compresses tokens into learned slices, applies slice attention and broadcasts the results back to the tokens; it also has a direct boundary cross-attention route. B2 and B3 are reduced GNOT-style and Transolver-style hosts, respectively (Hao et al., 2023; Wu et al., 2024).

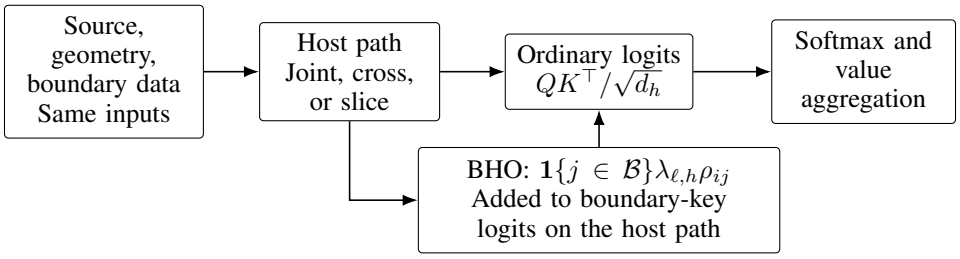


Figure 1: **Information, access, and the score intervention.** Both arms receive the same features. BHO modifies logits on the existing boundary path, preserving ordinary attention and value aggregation. At $\lambda = 0$, both arms have the same predictor. Here $\rho_{ij} = (q_i - k_j)^\top D_j(q_i - k_j)$. The residual acts on all query rows in the primary joint-host and OLBC bc1 comparisons, and on physical-interior queries in the separate BENO sign comparisons.

2.3 Score structure and attention support

Expanding the primary residual separates two self terms and a cross term:

$$\rho_{ij} = q_i^\top D_j q_i - 2q_i^\top D_j k_j + k_j^\top D_j k_j. \tag{3}$$

The cross-only control retains the middle term with the same conditioner; the full residual also includes both self terms. For key-invariant D, the query-self term is common to boundary keys. It shifts their logits relative to grid keys under joint attention and can change the total mass assigned to the boundary group. Under boundary-only softmax, a shift common to all participating keys cancels. The key-self and cross terms can still change selection within that group. With local D_j, the query-self term can also vary across keys and need not cancel. This algebra describes boundary-group reweighting and within-boundary selection at a fixed attention state (Appendix C).

2.4 Comparison conventions

Held-out BENO evaluations are distinguished from development-only ablations; OLBC populations are evaluated on their official test sets (Sections 3.3 and 4.2). Here "held-out" denotes exclusion from gradient training, not a previously unused evaluation population. Model comparisons are paired by training seed or subset/initialization block where applicable. Detailed data splits, normalization, initialization, estimators, confidence intervals, and test families are given in Appendices A–B, D, and E.

3 WHAT BOUNDARY-SCORE STRUCTURE CONTRIBUTES

The score intervention is evaluated first against an information-matched BENO host and then against specified simpler scores. On the OLBC circle, separately trained models are compared using solution accuracy and finite boundary response.

3.1 BENEFIT BEYOND SUPPLIED INFORMATION

Held-out evaluations use models trained for 300 epochs on 100 four-corner examples to assess source-shape prediction and transfer to five target shapes without fine-tuning. Separate development runs compare 100 examples with the full 800-example training pool in this BENO split under matched update budgets. Here N_{tr} counts gradient-training examples and "low-data" denotes the 100-example setting. Input and target scales use the 800-example training pool (Appendix A).

In the held-out evaluations, mean source-shape relative-L_2 error falls from 0.4574 to 0.3939, a 13.87% relative improvement. Transfer error decreases by 8.27%, with the five target-shape collections weighted equally (Table 1; Appendix A.5). These final-epoch results show a learning benefit beyond supplied information.

In the separate 3,900-update comparison, BHO reduces development error by 8.99% at $N_{\mathrm{tr}} = 100$, while the interval at $N_{\mathrm{tr}} = 800$ includes both benefit and harm (Table 1). The difference in BHO's relative effect between training sizes also remains unresolved (Appendix B.1, Table 9).

Table 1: **BHO benefit beyond the same supplied information.** Errors are arithmetic means $\pm$ standard deviations (SD) across eight paired seeds. Held-out models are trained with $N_{\mathrm{tr}} = 100$ for 300 epochs; each evaluation favors BHO in 8/8 pairs. Held-out changes use ratios of arithmetic seed means, with post-hoc paired-seed bootstrap 95% confidence intervals (CIs). The p-values are from two-sided paired-t tests with Holm adjustment over the three contrasts. The separate development models use 3,900 updates; their changes use paired log ratios and simultaneous 95% intervals across the two size-specific contrasts and their difference. Negative changes favor BHO.

Held-out BENO: $N_{\mathrm{tr}} = 100$, final epoch

Evaluation	Feature-only host	BHO	Change [95% CI]	Holm p
Source shape	0.4574 ± 0.0302	0.3939 ± 0.0210	-13.87% $[-17.76, -10.13]$	0.00116
Cross-shape transfer	0.4799 ± 0.0249	0.4402 ± 0.0132	-8.27% $[-11.79, -4.66]$	0.00536

Development BENO: common update budget

N_{tr}	Feature-only host	BHO	Change [simultaneous CI]
100	0.4285 ± 0.0550	0.3896 ± 0.0469	-8.99% $[-13.11, -4.67]$
800	0.1021 ± 0.0224	0.1102 ± 0.0094	$+9.94\%$ $[-14.54, 41.44]$

A separate 15,000-update schedule tests whether the advantage persists with more training. The reduction at $N_{\mathrm{tr}} = 100$ is 11.45%; neither the effect at $N_{\mathrm{tr}} = 800$ nor the change in the low-data BHO/host ratio between schedules is resolved (Appendix B.2, Table 11). The longer low-data runs fit the training data more closely while development error rises, with BHO error still lower than the host's (Appendix B.2, Figure 5). The advantage thus persists across the two tested horizons, without establishing convergence.

Separate development-set sensitivity analyses also show improvement with subset-only normalization and independently sampled training subsets (Appendices A.3 and A.4).

3.2 SIMPLER SCORE CONTROLS IN THE BENO HOST

The BENO score controls use the same 100-example, 3,900-update setting. HOST denotes the feature-only control. GATE introduces a scalar bias per layer and head that reweights the boundary group without changing its conditional within-boundary weights at a fixed attention state.

BILINEAR retains only $-2\lambda q_i^\top D_j k_j$, with BHO's local conditioner and parameter count; BHO also retains both self terms (Equation (3)). Paired arms share the initial host predictor, training subset, data order and budget (Appendix B.1).

On the BENO development set at $N_{\mathrm{tr}} = 100$, the full quadratic reduces error by 6.35% relative to the conditioned cross-only score. A secondary comparison estimates a 7.68% reduction relative to the static group bias (unadjusted paired-t $p = 0.0002$). Both comparisons reuse the BHO runs from the primary BHO–host comparison (Table 2; Appendix B.1). Figure 2 shows paired seed errors and BHO/control changes, separating pointwise from family-simultaneous intervals.

The BHO–BILINEAR contrast is significant after Holm correction within the five-contrast family; BHO–GATE is secondary and unadjusted. As in the other full-versus-cross-only comparisons, initial predictions are matched, but residual scale and gate sensitivity are not. The contrast therefore compares learning parameterizations rather than isolating the self terms (Appendix E).

Table 2: **Simple score controls in the BENO host.** Development error, mean ± SD across eight paired seeds at $N_{\mathrm{tr}} = 100$ and 3,900 updates. Training uses rows 0–99; normalizers and the target scale are fitted on the 800-example pool (rows 0–799). These four arms are drawn from a nine-arm matched development study; correction families are specified in Appendix B.1. HOST and BHO use the same endpoints as Table 1. Percentage changes use paired log ratios (Appendix A.2), not ratios of these arithmetic means.

Arm	Added parameters	Development error
HOST	0	0.4285 ± 0.0550
GATE (static group bias)	32	0.4224 ± 0.0543
BILINEAR (conditioned cross term)	2,080	0.4166 ± 0.0539
BHO	2,080	0.3896 ± 0.0469

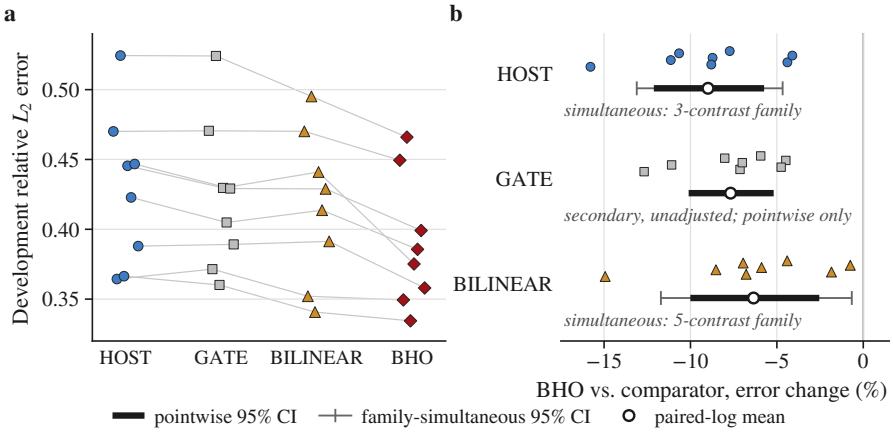


Figure 2: **Paired BENO score-control comparisons.** GATE denotes the static boundary-group bias; BILINEAR denotes the conditioner-matched cross-only score. Left: development relative-L_2 endpoints for all eight paired seeds at $N_{\mathrm{tr}} = 100$ and 3,900 updates. Right: points show seed-level percentage error changes; open circles show BHO/control effects estimated from paired log ratios. Thick intervals are pointwise 95% CIs; thin intervals are family-simultaneous 95% CIs (three-contrast HOST and five-contrast BILINEAR families). The GATE contrast is secondary and unadjusted, with no simultaneous interval. The HOST and BHO endpoints are those in Table 1, not an independent replication.

3.3 Score controls and finite boundary response

The OLBC circle comparison separately trains the same four arms in a reduced-resolution B1-style joint-attention host. Geometry and source are fixed, while Dirichlet values vary. Eight paired

subset/initialization blocks use $N_{\mathrm{tr}} = 100$ and 2,100 updates, followed by evaluation on all 256 official test examples. This test population had been used in earlier exploratory analyses. The primary metric is solution relative-L_2 error. A secondary metric measures the response to boundary changes by comparing $\hat{u}(g_2) - \hat{u}(g_1)$ with $u(g_2) - u(g_1)$ on 128 prespecified disjoint pairs. For prediction error $e_u(g) = \hat{u}(g) - u(g)$, the response error is $e_u(g_2) - e_u(g_1)$: any component independent of g cancels. This measures how prediction error varies with boundary data (Appendix D).

On the primary solution metric, BHO reduces error by 10.93% relative to BILINEAR, while its differences from HOST and GATE remain unresolved under the three-comparison Holm correction (Table 3; Appendix D, Table 13). On the secondary response metric, all three contrasts favor BHO after Holm correction within the separate secondary family, including a 9.68% reduction relative to HOST. Both metrics use the same predictions.

Table 3: **Score structure beyond the same information and route.** OLBC bc1, $N_{\mathrm{tr}} = 100$, official test set. Entries are geometric means across eight paired subset/initialization blocks. Solution error averages per-example relative-L_2 errors; the secondary response error pools finite differences over 128 fixed pairs. Only the BHO–BILINEAR contrast is resolved in the primary Holm family; all three BHO/control response contrasts are resolved in their separate secondary family. Relative changes are $100(G_{\mathrm{BHO}}/G_{\mathrm{control}} - 1)$, using the unrounded geometric means G (Appendix A.2, Equation (4)).

Arm	Boundary-score intervention	Solution	Boundary response
HOST	None	0.0663	0.0996
GATE	Group logit bias	0.0669	0.1007
BILINEAR	Conditioned cross term	0.0694	0.1042
BHO	Full conditioned quadratic	0.0618	0.0900

Figure 3 illustrates a boundary-induced solution change and compares response errors across matched training blocks. Aggregate response comparisons use all 128 prespecified pairs in each block; a post-hoc all-pairs analysis gives effects within 0.5 percentage points of those from the prespecified analysis (Appendix D.1).

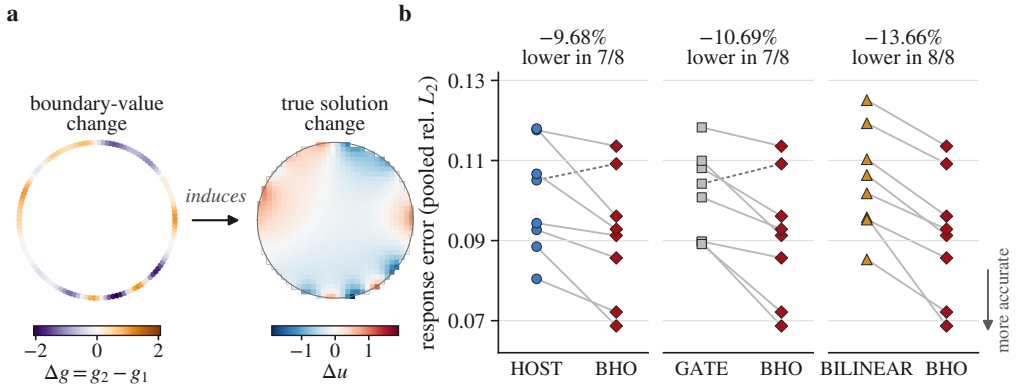


Figure 3: **Predicting the solution change induced by boundary data.** OLBC bc1, fixed geometry and source. (a) Boundary-value change Δg and true solution change Δu for the first prespecified pair in the lowest-seed block (physical units; separate color scales). (b) Response relative-L_2 error pooled over 128 pairs per block. Each line connects a control and BHO within one of eight training blocks; dashed lines indicate higher BHO error. Headers show paired-log percentage changes and counts of blocks with lower BHO error. Pointwise 95% intervals and Holm-adjusted tests are in Appendix D, Table 13.

To assess task difficulty, an affine-ridge reference uses the same boundary observations and training subsets. It achieves a mean test relative-L_2 error of 3.43×10^{-8} (Appendix D, Table 14).

Together, these results show host-relative benefit in the tested low-data BENO setting and a full-versus-cross-only advantage in BENO and on the circle. The circle's primary contrasts against HOST and GATE remain unresolved. Its affine-ridge result also shows that this neural comparison does not establish a need for a neural predictor.

4 SCOPE ACROSS HOSTS AND POPULATIONS

The joint-attention results motivate comparisons across access designs and populations. Within each host and population, the compared arms receive the same boundary information and use the same access path.

4.1 BOUNDARY-ACCESS PATHS

The BENO comparisons use the full training pool ($N_{\mathrm{tr}} = 800$) in the cross-attention (B2) and slice (B3) hosts, both with boundary-only routes (Section 2.2). BHO reduces B2 development error by 47.2% (protocol and post-hoc uncertainty in Appendix E). The B3 comparison matches architecture and initialization while training or freezing residual gates. Its estimated change is $+0.67\%$, with an interval compatible with benefit or harm (Table 4). These are complete architectures that differ beyond the boundary route.

Table 4: **Score interventions on existing boundary-access paths.** BENO development at $N_{\mathrm{tr}} = 800$, with boundary inputs and access matched within each host. For B2, errors are arithmetic seed means; the change uses their unrounded ratio, with a post-hoc paired-seed bootstrap 95% CI. The B3 comparison also matches route initialization and uses a pointwise 95% paired-log interval.

Host / boundary path	Host	BHO	Error change [95% CI]
B2 / cross-attention	0.2342	0.1236	-47.2% $[-52.2, -41.9]$
B3 / direct route	0.1445	0.1452	$+0.67\%$ $[-7.56, 9.64]$

4.2 SCORE CONTROLS ON A NATIVE BOUNDARY-AWARE HOST

The native comparison tests whether BHO improves on the released OLBC extended Geometry Aware Operator Transformer (XGAOT), which already uses a learned boundary extension. Both Poisson–Boomerang populations have fixed non-convex geometry: Mixed varies boundary data with a fixed source; Mixed+ varies both. All arms receive the same boundary-feature augmentation and are trained separately on each population.

BILINEAR and BHO modify the boundary-only route with conditioned cross-only and full quadratic scores. A uniform logit bias cancels on this route, so native GATE scales the boundary message. For each population, eight paired subset/initialization blocks use $N_{\mathrm{tr}} = 1,024$ and 24,000 updates. Host initialization, preprocessing, batch order and checkpoint selection are matched within each block. The primary metric is mean per-example untrimmed relative-L_2 error, computed over all valid native nodes, including boundary nodes, without grid interpolation. Official test evaluation uses validation-selected weights (Appendix D.2).

Neither comparison establishes BHO's superiority over HOST or the message gate under its three-contrast Holm family; this does not establish equivalence. BHO–BILINEAR is resolved only on Mixed, with a 2.89% reduction; no primary Mixed+ contrast is resolved (Table 5). These comparisons use the released Robin host channel convention; a metadata-aligned host encoding has not been evaluated (Appendix D.2).

The full quadratic thus also outperforms the cross-only score on Mixed, while its advantage over the native host remains unresolved. On Steel–Circle^H elasticity, BHO has higher primary error than HOST and GATE, with its BILINEAR contrast unresolved (Appendix F).

Table 5: **Score controls on native Poisson–Boomerang populations.** Primary mean per-example untrimmed relative-L_2 error on each official test set, using validation-selected weights. Changes use paired log ratios across eight blocks. Intervals are pointwise 95%. Holm adjustment applies separately to the three contrasts in each population; significance is assessed at the 5% level.

Population	Contrast	Change [pointwise 95% CI]	Holm p
Mixed	BHO/HOST	-1.69% $[-3.21, -0.14]$	0.0741
	BHO/GATE	-1.06% $[-3.49, +1.43]$	0.343
	BHO/BILINEAR	-2.89% $[-4.65, -1.09]$	0.0209
Mixed+	BHO/HOST	$+0.60\%$ $[-1.46, +2.70]$	1.00
	BHO/GATE	$+1.56\%$ $[-0.93, +4.11]$	0.551
	BHO/BILINEAR	$+0.52\%$ $[-1.75, +2.84]$	1.00

5 Parameterization-dependent constraint costs

Q2 asks whether within-head sign mixing can be removed without losing accuracy, and whether the answer depends on separate or shared residual projections. This separate study uses a constant conditioner input and physical-interior queries; Q1 uses the host's directed Q, K, local D_j and all query rows. In both studies, boundary features remain host inputs.

5.1 Projection and spectrum constraints

The directed residual reuses the host's separate Q, K; the shared residual uses $(P_h h_i, P_h h_j)$ with a residual-only projection of dimension $r = 16$, while ordinary attention retains its own Q, K. Both the mixed (MIX) $\tanh(We)$ and homogeneous (HOM) $|\tanh(We)|$ parameterizations retain freely signed head gates. The restriction removes within-head coordinate sign mixing without imposing a single sign on the whole model. MIX and HOM are compared within each family across eight paired seeds at $N_{\text{tr}} = 100$, with matched parameter counts, raw adapter and host initialization, and data order (Appendix A.1). Each HOM/MIX error ratio measures a restriction cost. The interaction reports the percentage change in the shared-family ratio relative to the directed-family ratio.

Homogeneity increases development error by 23.19% in the shared family and 0.71% in the directed family, with an interaction of 22.32% (Figure 4; Appendix A.2, Table 8). No cost is detected in the directed family; this does not establish equivalence. Both directed variants improve on the feature-only host.

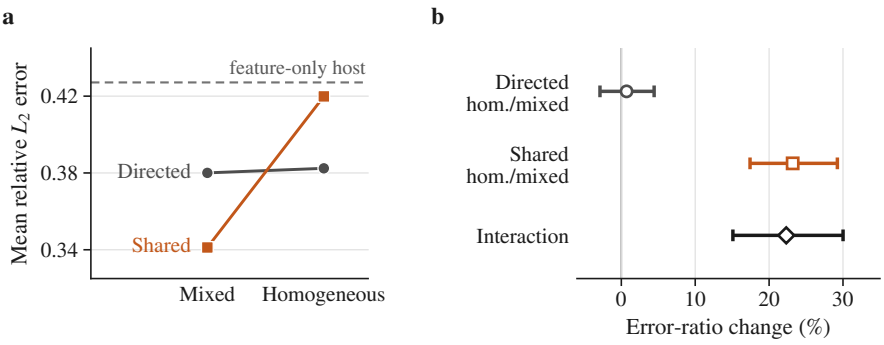


Figure 4: **The cost of removing sign mixing depends on parameterization.** Left: mean BENO development errors at $N_{\text{tr}} = 100$, with the common feature-only host error shown by the dashed line. Right: homogeneous/mixed costs and their interaction, with pointwise 95% paired-log intervals across eight confirmatory seeds. Spectrum comparisons are matched within each family; shared and directed families differ in capacity and coupling to ordinary attention.

The shared-family penalty also persists at 3,900 matched updates with fixed initial signs in a separate study (Appendix B.3).

For free hidden vectors and projections at constant D, positive-definite D permits every rank-at-most-r isolated directed cross-coupling $Q^\top DK$, whereas shared $P^\top DP$ is semidefinite for homogeneous D, up to the head gate's sign. The full directed residual remains restricted by coupled self terms (Appendix C). These standard rank and inertia facts characterize the specified algebraic restrictions; related attention parameterizations and weighted inner-product similarities have been studied previously (Tsai et al., 2019; Kim et al., 2019). The shared and directed families differ in capacity and coupling to ordinary attention, so the empirical interaction compares restriction costs across these complete parameterizations.

5.2 Fixed versus learned coordinate signs

The shared-family penalty raises a further question: must the mixed signs be learned? A separate shared-model comparison fixes their initial assignment while learning magnitudes; another arm fixes both signs and magnitudes. Fixing signs alone gives a one-sided upper bound of 0.43% on excess error relative to free signs, within the 3% non-inferiority margin. Fixing both signs and magnitudes gives an upper bound of 3.85% and does not establish non-inferiority (Appendix B.4, Table 12). A fixed mixed pattern therefore sufficed for this shared model within the stated margin; the result concerns learning its assignment while retaining sign mixing. These parameterization-dependent restriction costs do not identify the cause of the primary host-relative gains.

6 Discussion and limitations

Boundary-score parameterization can affect learning even when information and access are matched. BHO improves source-shape and cross-shape accuracy in low-data BENO and retains its advantage under two matched development budgets.

The simpler-score and host comparisons support different conclusions. The full quadratic outperforms conditioner-matched cross-only scores in BENO, the OLBC circle and native Mixed, but the distinction is unresolved on Mixed+. Primary comparisons against HOST and GATE remain unresolved on the circle and both native populations, while Steel–Circle^H elasticity favors HOST and GATE (Appendix F). Improvement over a specified simpler score thus does not establish a benefit over the host across populations. The native results use the released Robin host convention; a metadata-aligned host encoding has not been evaluated.

The circle's affine-ridge result also limits the task-difficulty interpretation: its neural comparisons assess learning parameterizations, not a need for a neural predictor on this fixed-source population.

The separate constraint study shows that simplification costs depend on the parameterization. With constant conditioning and physical-interior queries, sign homogeneity is costly in the shared family and has no detected cost in the directed family. The interaction compares restriction costs across families that differ in capacity and attention coupling. Within the shared model, fixed initial mixed signs are non-inferior within the 3% margin: access to a mixed pattern suffices without learning its assignment under that criterion.

These conclusions concern the tested hosts, populations and finite training budgets. The source-shape relative-L_2 error of 0.3939 shows a learning improvement but does not establish practical solver accuracy (Table 1); the training-size interaction remains unresolved.

AI use statement

Language-model assistants supported hypothesis and protocol development, implementation and code inspection, mathematical analysis, interpretation, and manuscript editing. Candidate explanations were evaluated through analyses and experiments. The authors are responsible for all claims, code, and results.

Ethics statement

This study uses simulated and public PDE benchmarks. Learned surrogates can produce inaccurate scientific predictions; the findings are confined to the evaluated numerical settings.

Reproducibility statement

Appendices A–B, D, and E describe the data splits, preprocessing, model configurations, training schedules, and statistical estimators; Appendix C gives proofs. Paired seed- or block-level endpoints and figure data accompany the paper (Appendix E). An anonymous repository for this work is available at `https://anonymous.4open.science/r/bho-iclr2027`.

References

Sumanth Kumar Boya and Deepak N. Subramani. PINTO: Physics-informed transformer neural operator for learning generalized solutions of partial differential equations for any initial and boundary condition. *Computer Physics Communications*, 315:109702, 2025. doi: 10.1016/j.cpc.2025.109702. URL `https://doi.org/10.1016/j.cpc.2025.109702`.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 2023. URL `https://proceedings.mlr.press/v202/hao23c.html`.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL `https://www.jstor.org/stable/4615733`.

Masanobu Horie and Naoto Mitsume. Physics-embedded neural networks: Graph neural PDE solvers with mixed boundary conditions. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1687. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/93476ae409ae3246e22a9d4b931f84ed-Abstract-Conference.html`.

Geewook Kim, Akifumi Okuno, Kazuki Fukui, and Hidetoshi Shimodaira. Representation learning with weighted inner product for universal approximation of general similarities. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 5031–5038, 2019. doi: 10.24963/ijcai.2019/699. URL `https://www.ijcai.org/proceedings/2019/699`.

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. URL `https://jmlr.org/papers/v24/21-1524.html`.

Ziyuan Liu, Yuhang Wu, Daniel Zhengyu Huang, Hong Zhang, Xu Qian, and Songhe Song. SPFNO: Spectral operator learning for PDEs with Dirichlet and Neumann boundary conditions. *arXiv preprint arXiv:2312.06980*, 2023. URL `https://arxiv.org/abs/2312.06980`.

Sepehr Mousavi, Siddhartha Mishra, and Laura De Lorenzis. Imposing boundary conditions on neural operators via learned function extensions. In *International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=gnzNB7GFFo`.

Nadim Saad, Gaurav Gupta, Shima Alizadeh, and Danielle C. Maddix. Guiding continuous operator learning through physics-based boundary constraints. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=gfWNItGOES6`.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2074. URL `https://aclanthology.org/N18-2074/`.

James Joseph Sylvester. XIX. A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 4(23):138–142, 1852. doi: 10.1080/14786445208647087. URL https://doi.org/10.1080/14786445208647087.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4344–4353. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1443. URL https://aclanthology.org/D19-1443/.

Haixin Wang, Jiaxin Li, Anubhav Dwivedi, Kentaro Hara, and Tailin Wu. BENO: Boundary-embedded neural operators for elliptic PDEs. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/218ca0d92e6ed8f9db00621e103dc70c-Abstract-Conference.html.

Shizheng Wen, Arsh Kumbhat, Levi E. Lingsch, Sepehr Mousavi, Yizhou Zhao, Praveen Chandrashekar, and Siddhartha Mishra. Geometry aware operator transformer as an efficient and accurate neural surrogate for PDEs on arbitrary domains. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5197. URL https://papers.nips.cc/paper_files/paper/2025/hash/e45a448dfa778f6d62729a7bc8633c06-Abstract-Conference.html.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53681–53705. PMLR, 2024. URL https://proceedings.mlr.press/v235/wu24r.html.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, volume 34, pp. 28877–28888, 2021. URL https://papers.neurips.cc/paper_files/paper/2021/hash/f1c1592588411002af340cbaedd6fc33-Abstract.html.

A Experimental design and statistical analysis

The BENO experiments use the released arrays (Wang et al., 2024). The Neumann label identifies the released population; its boundary discretization, compatibility treatment, and additive-constant convention are not fully specified by the benchmark documentation. The comparisons evaluate predictions of the released solutions, without spatial centering or an imposed gauge. Unless stated otherwise, training uses the first N_{tr} four-corner examples from rows 0–799; development error is evaluated on rows 800–899, and held-out evaluation uses rows 900–999. Throughout the appendices, N abbreviates N_{tr}. Table 6 summarizes normalization, paired units and the purpose of each comparison.

For both matched-budget comparisons, normalizers and a target scale are fitted on the 800-example training pool. Thus N_{tr} counts gradient-training examples rather than preprocessing examples. Appendix A.3 compares pool-fitted and subset-only normalization; Appendix D specifies train-only normalization for OLBC bc1.

A.1 Factors and matched comparisons

Q1's primary comparisons use directed host Q, K, local D_j and all query rows, comparing the full quadratic against the same-information host and specified simpler scores. Q2 instead measures homogeneous/mixed restriction costs within the constant-conditioned, physical-interior families below. In both settings the full quadratic is added to the scores on the host's ordinary attention path.

Table 6: **Study design and data conventions.** BENO evaluation is on the four-corner development block unless labeled held-out. "Pool" denotes normalization on the 800-example training pool; "subset" denotes the examples used for optimization.

Study	Independent unit	Normalizers	Purpose
Equal-update BHO–host	Seeds 47000–47007	Pool	Primary comparison at $N = 100, 800$
Longer matched budget	Same seeds 47000–47007	Pool	Budget robustness at $N = 100, 800$
BENO simple-score controls	Same seeds 47000–47007	Pool	Static bias and conditioned cross term
Held-out BENO, $N = 100$	Seeds 0–7	Pool	Source-shape and transfer tests
Subset normalization	8 paired seeds	Subset	Normalization sensitivity
Projection–spectrum	Seeds 8–15	Pool	Projection–spectrum interaction
Fixed-sign study	Seeds 16–23	Pool	Sign-learning ablation
Matched B3 residual comparison	Seeds 100–107	Matched across arms	Residual on a matched direct route
Equal-update spectrum control	8 subset/initialization pairs	Nested 100-example subset	Data and update moderation
OLBC bc1 score controls	8 subset/init blocks, seeds 53000–53007	Subset	Solution and boundary-response tests
Boomerang Mixed	8 subset/init blocks, seeds 62000–62007	Matched across arms	Native mixed-boundary score controls
Boomerang Mixed+	8 subset/init blocks, seeds 63000–63007	Matched across arms	Score controls with variable source

The projection–spectrum comparison crosses directed versus shared residual projections with mixed versus homogeneous diagonals. Its general score form is

$$s_{ij} = \frac{q_i^\top k_j}{\sqrt{d_h}} + m_i \mathbf{1}\{j \in \mathcal{B}\} \lambda_{\ell,h} (z_i - z_j)^\top D_j (z_i - z_j),$$

where m_i selects the query support. Layer/head indices on projections and conditioners are suppressed. The directed residual uses $z_i = Qh_i$, $z_j = Kh_j$; the shared residual uses $z_i = P_h h_i$, $z_j = P_h h_j$ with a residual-only projection of latent dimension 16. Ordinary attention continues to use its own Q, K. Mixed and homogeneous spectra use $\mathrm{diag}(\tanh(W_h e))$ and $\mathrm{diag}(|\tanh(W_h e)|)$, respectively. Both retain a freely signed head gate, so the restriction removes within-head coordinate sign mixing rather than fixing the whole head's sign.

The eight-seed confirmatory comparison uses physical-interior queries and constant conditioning $e_{\mathrm{const}} = (0, 0, -1, 0)^\top$ in the Neumann-labeled BENO setting. Estimated normals, boundary type, and varying boundary values remain ordinary token inputs. The arms in each within-family comparison share the host initialization, raw W_0, data order, and P_0 where applicable. Directed and shared models have 548,769 and 614,305 parameters, respectively. Spectrum comparisons are parameter-count matched within each family; comparisons across families also change capacity and coupling to ordinary attention.

A.2 Estimates and statistical tests

For endpoints $e_{u,s}, e_{v,s} > 0$ from paired unit s (a training seed or subset/initialization block), let $z_s = \log(e_{u,s}/e_{v,s})$. Unless otherwise specified, the reported paired-log change, with a pointwise

95% confidence interval, is

$$\Delta_{u/v} = 100[\exp(\bar{z}) - 1], \qquad \bar{z} = \frac{1}{n}\sum_{s=1}^{n} z_s. \tag{4}$$

Here n is the number of paired units, and negative changes favor the numerator. With sample standard deviation s_z of the log ratios, exponentiating $\bar{z} \pm t_{0.975,n-1} s_z/\sqrt{n}$ gives a pointwise 95% interval for the error ratio. Applying $100(r-1)$ to each bound r gives the interval for the percentage change in Equation (4).

In Tables 7–8, k/n counts paired units favoring the stated direction. The unit is the training seed unless a row explicitly counts independently sampled subsets. Fixed-dataset intervals, including those for the equal-update BHO–host comparison, quantify training variation conditional on the common training and evaluation sets. The shared-spectrum equal-update study also resamples the training subset in each replicate.

Exact Wilcoxon signed-rank tests use the signs and absolute ranks of paired differences; exact sign-flip tests use study-specific statistics. Both require independent paired units and their respective sign-symmetry assumptions. Numerical values are rounded for display.

Where specified, Holm correction adjusts p-values within each study's test family (Holm, 1979). The equal-update BHO–host and shared-spectrum studies report Bonferroni simultaneous confidence intervals across three and four contrasts, respectively (Appendix B). The longer-budget BHO–host comparisons use separate prespecified adjustments for endpoint and between-schedule tests (Appendix B.2). The OLBC bc1 solution and response endpoints use separate three-contrast Holm families with pointwise intervals (Appendix D). Each native Boomerang population has its own three-contrast Holm family, also reported with pointwise intervals (Appendix D.2). Exact sign-flip tests for the primary equal-update comparison are sensitivity analyses. Non-inferiority requires the one-sided upper bound to fall within the prespecified margin.

Table 7: **Tests supporting the performance comparisons.** Changes are versus the stated control. "Favors BHO" counts negative error changes; subset resampling counts subsets. Estimators and test families follow the respective studies. Primary equal-update tests are in Table 9; longer-budget tests are in Table 11. The matched slice-route interval is $[-7.56, 9.64]\%$.

Comparison	Change	Favors BHO	Test
Subset-only normalizers, $N = 100$	-12.92%	8/8	Exact Wilcoxon, $p = 0.0078$
Independent training subsets	-8.4%	7/8 subsets	Paired-difference t, $p = 0.003$
Held-out BENO, source shape	-13.87%	8/8	Holm-adjusted $p = 0.00116$
Held-out BENO, transfer	-8.27%	8/8	Holm-adjusted $p = 0.00536$
Cross-attention host, $N = 800$	-47.2%	8/8	Exact Wilcoxon, $p = 0.0078$
Matched slice-route residual	$+0.67\%$	3/8	Exact Wilcoxon, $p = 0.9453$

Table 8: **Confirmatory sign-restriction contrasts.** BENO $N = 100$, constant conditioner, interior support, seeds 8–15. Counts indicate a positive cost or interaction. Intervals are pointwise 95% paired-log CIs. The separate full-query check gives a 24.94% interaction. At $N = 200$, the 300-epoch projection–spectrum comparison gives an interaction of 2.85% (95% CI $[-6.34, 12.94]\%$).

Contrast	Change [95% CI]	Positive pairs	Exact Wilcoxon p
Directed HOM/MIX	$+0.71\%$ $[-2.89, 4.44]$	3/8	0.9453
Shared HOM/MIX	$+23.19\%$ $[17.42, 29.24]$	8/8	0.0078
Interaction	$+22.32\%$ $[15.10, 30.01]$	8/8	0.0078

A.3 Normalization sensitivity

This sensitivity analysis compares preprocessing fitted on the full training pool with preprocessing fitted only on the optimization subset. The respective control/BHO errors are 0.4390/0.3898 and 0.4463/0.3886; the reported within-condition gains are 11.16% and 12.92%. Separately, a paired comparison between normalization choices gives a mean seed-wise change of $+1.67$ percentage points in BHO gain with subset-only normalization (95% CI $[-1.64, 4.98]$). Thus the comparison does not establish that subset-only normalization increases the gain.

A.4 Training-subset sensitivity

Resampling the training examples tests whether the low-data benefit is specific to the initial subset. Across eight independently sampled 100-example subsets, averaging two initializations within each subset, BHO reduces mean error by 8.4% and is favored in seven of eight subsets (paired-difference t, $p = 0.003$).

A.5 Held-out BENO evaluations

The feature-only host and locally conditioned signed-$\tanh$ BHO are trained on 100 examples of the Neumann-labeled four-corner population and evaluated on a disjoint 100-example source-shape test block. For transfer, the same trained weights are applied without fine-tuning to five BENO target collections, whose 100-example test means are weighted equally. Each example's relative L_2 error is evaluated on the physical-interior subset of the 1,024 grid positions.

Training uses 300 epochs with batch size 16, giving 2,100 updates with a four-example final batch in each epoch. AdamW uses a peak learning rate of 10^{-3}, weight decay of 10^{-5}, and a OneCycle schedule; the final-epoch checkpoint is evaluated without development-set model selection. Source, boundary-value, and target scales are fitted on the 800-example training pool and applied unchanged to all evaluation shapes. Within each paired run, the two arms share the host initialization and batch order. BHO's conditioner uses an arm-specific draw from $\mathcal{N}(0, 1/4)$ (variance $1/4$), and its zero gates give the same initial predictor as the host.

For seed errors $E_{\mathrm{BHO},s}$ and $E_{\mathrm{host},s}$, the held-out change is $100\,\overline{(E_{\mathrm{BHO}} - E_{\mathrm{host}})}/\overline{E_{\mathrm{host}}}$, rather than the paired-log estimator. The prespecified primary family comprises source-shape BHO versus host (F1), a variant omitting normal/type token channels while retaining them in the conditioner versus host on the source shape (F2), and cross-shape BHO versus host (F3). Holm adjustment uses all three two-sided paired-t tests on the untransformed seed errors. The adjusted p-values for the source-shape and transfer evaluations are 0.00116 and 0.00536, respectively, with BHO favored in 8/8 pairs in each case; both exact two-sided Wilcoxon checks give $p = 0.0078$. F2 reduces source-shape error by 20.85% (post-hoc paired-seed bootstrap 95% CI $[-24.76, -16.69]\%$; Holm $p = 0.000310$; 8/8 pairs). This contrast changes both the ordinary token channels and the score.

For the displayed arithmetic-mean-ratio effects, post-hoc paired-seed bootstrap 95% CIs are computed using 200,000 resamples of the eight seed pairs, with replacement and random seed 20260924. Each resample preserves pairing between each compared arm and HOST and evaluates $100(\overline{E}_{\mathrm{arm}}/\overline{E}_{\mathrm{host}} - 1)$; the interval is bounded by the 2.5th and 97.5th percentiles of the bootstrap distribution. The transfer endpoint averages the five target shapes within each seed before resam-

pling. These intervals are conditional on the fixed datasets and are separate from the prespecified F1–F3 tests and Holm adjustment.

The held-out protocol and primary contrasts were fixed before test evaluation. These rows are excluded from gradient training; later reuse of the same populations is not treated as independent confirmation. The other BENO ablations reported here use development rows 800–899. OLBC data use is specified in Appendix D.

B EQUAL-UPDATE COMPARISONS

B.1 PRIMARY BHO VERSUS FEATURE-ONLY HOST

This comparison tests whether the directed residual improves its information-matched host under one common training schedule. The two arms are the locally conditioned, signed-$\tanh$ BHO from Section 2.2 and the feature-only host, evaluated at $N = 100$ and 800. Both use the same estimated normals, token features, and joint attention over 1,152 keys. BHO acts on boundary keys for every query; its gates and conditioner weights are trainable.

The 32 runs combine eight seeds (47000–47007), two arms and two training sizes. Within each seed, all four conditions share the host initialization. The two BHO training-size conditions reuse the same conditioner tensor, whose entries are drawn independently from $\mathcal{N}(0, 1/4)$ (variance $1/4$), and both initialize their gates at zero.

The training sets are fixed, nested subsets comprising rows 0–99 and 0–799 of the Neumann-labeled four-corner population. Input normalizers and the target scale are fitted once on rows 0–799 and reused across all conditions. Evaluation averages the error in Equation (1) over development rows 800–899. The inferential unit is the paired seed block; training and evaluation examples are fixed.

All models use a common 3,900-update OneCycle schedule. Batches of 16 are formed from concatenated, independently shuffled permutations of the training set, with remaining examples carried across permutation boundaries. Each run uses 62,400 example presentations. Within a training size and seed, BHO and its host receive identical batch sequences. The primary endpoint is update 3,900; update 2,100 is a secondary checkpoint of the same schedule. Appendix E specifies optimization and precision.

Primary contrasts. For mean development error $e_{N,a,s,t}$ of arm a, seed s, and checkpoint t, define

$$z_{N,s,t} = \log \frac{e_{N,\mathrm{BHO},s,t}}{e_{N,\mathrm{host},s,t}}.$$

The three prespecified primary contrasts are

$$G_{100,s} = z_{100,s,3,900}, \qquad G_{800,s} = z_{800,s,3,900}, \qquad M_s = G_{800,s} - G_{100,s}.$$

Negative G_{100} and G_{800} favor BHO; positive M indicates a larger BHO/host ratio at $N = 800$. Contrasts are formed within seed. Reported ratios are obtained by exponentiating mean log contrasts, with Bonferroni-adjusted simultaneous 95% intervals over the three contrasts. Table 9 also reports unadjusted paired-t p-values and exact sign-flip sensitivity tests over all 2^8 sign assignments.

Table 9: **Primary BHO–host contrasts at 3,900 updates.** Eight paired seeds on fixed datasets. Ratios are obtained by exponentiating mean log contrasts; intervals provide Bonferroni-adjusted simultaneous 95% coverage over the three primary contrasts. p-values are unadjusted; sign-flip tests are sensitivity analyses.

Contrast	Ratio [simultaneous CI]	Paired-t p	Sign-flip p
G_{100}	$0.9101\ [0.8689, 0.9533]$	0.00038	0.0078125
G_{800}	$1.0994\ [0.8546, 1.4144]$	0.278	0.3203125
M	$1.2080\ [0.9040, 1.6142]$	0.0809	0.09375

The simultaneous interval establishes a low-data improvement, with BHO favored in eight of eight pairs at $N = 100$. At $N = 800$, four of eight pairs favor BHO; neither its effect nor the difference between training sizes is resolved by the simultaneous intervals.

Intermediate-checkpoint analysis. The secondary contrasts are the two within-size effects and their difference at update 2,100, together with $U_{100,s} = z_{100,s,2,100} - z_{100,s,3,900}$ (Table 10). These use pointwise 95% intervals; both checkpoints belong to the same 3,900-update schedule.

Table 10: **Secondary contrasts within the common schedule.** Eight paired seeds; pointwise 95% paired-t intervals on the log scale, exponentiated as ratios.

Contrast	Ratio [pointwise 95% CI]	Paired-t p
$z_{100,2,100}$	$0.9225\ [0.8852, 0.9613]$	0.00241
$z_{800,2,100}$	$0.9899\ [0.8608, 1.1384]$	0.869
$z_{800,2,100} - z_{100,2,100}$	$1.0731\ [0.9156, 1.2578]$	0.328
U_{100}	$1.0135\ [0.9983, 1.0290]$	0.0745

Matched simple-score controls. Table 2 shows four arms from a nine-arm development comparison; these four use the same full-feature B1 host. The score acts on all 1,152 query rows and 128 boundary keys. The static bias has one coefficient per layer and head; BILINEAR and BHO use $e_j = (n_x, n_y, \text{type}, g)$ with the same local conditioner $D_j = \text{diag}(\tanh(We_j))$. Their conditioner weights are initialized from the same draw and trained separately. Zero initial score coefficients give the same initial predictor, with matched host weights and batch streams. The comparison uses the same Neumann-labeled four-corner training rows 0–99, development rows 800–899, pool-fitted preprocessing, seeds 47000–47007, and terminal 3,900-update AdamW/OneCycle endpoint specified above. HOST and BHO use the primary equal-update endpoints and their three-contrast inference.

The BHO–BILINEAR contrast uses paired-t tests on log errors, with Holm-adjusted p-values and Bonferroni simultaneous 95% intervals over the full five-contrast primary family. The estimated change is -6.35% (simultaneous 95% CI $[-11.71, -0.67]\%$; Holm $p = 0.0297$). Its pointwise 95% interval is $[-10.01, -2.55]\%$. The BHO–static-bias comparison is secondary: the estimated change is -7.68% (pointwise 95% CI $[-10.11, -5.19]\%$; unadjusted paired-t $p = 0.0002$). Both contrasts favor BHO across all eight paired seeds. The accompanying analysis files provide paired endpoints and all five primary contrasts.

B.2 Longer matched-budget control

This control tests whether the low-data BHO–host advantage persists under a longer matched training budget. The directed, locally conditioned BHO and its full-feature host use the same paired cohort, seeds 47000–47007, and the same fixed training sets, pool-fitted preprocessing, and development rows as Appendix B.1. Held-out and transfer data are not used. Each run starts from its original initialization with a 15,000-update OneCycle schedule fixed before training; the batch-stream prefix matches the short schedule. The primary endpoint is update 15,000, without selecting the checkpoint with the lowest development error. Because the two OneCycle schedules have different horizons, their states at update 3,900 are not equivalent training endpoints. Optimization and precision otherwise follow Appendix E.

Endpoint and between-schedule comparisons each use a four-contrast family, with Holm-adjusted paired-t tests and Bonferroni simultaneous 95% confidence intervals (98.75% per contrast). The inferential unit is the paired seed on fixed training and development sets; all results in this control are development-only.

For the BHO–host contrast, the between-schedule change is $H_{N,s} = z^{\text{long}}_{N,s,15,000} - z^{\text{short}}_{N,s,3,900}$. Exponentiating its mean gives 0.9729 at $N = 100$ (simultaneous CI $[0.9440, 1.0027]$, Holm $p = 0.0755$) and 0.9294 at $N = 800$ ($[0.6879, 1.2558]$, Holm $p = 0.809$). Thus BHO improves low-data error under both tested budgets, without an established change in its relative effect.

Training trajectories and late-stage continuation. Each 15,000-update trajectory was continued for a prespecified 3,000 updates at a constant learning rate of 10^{-5} to assess late-stage behavior. The continuation used $\beta_1 = 0.9$ and retained Adam moments, β_2, weight decay, and ϵ, with the OneCycle scheduler no longer stepped. At $N = 100$, host/BHO training errors fall from 0.0261/0.0246

Table 11: **BHO–host contrasts under the longer matched budget.** Development errors at the prespecified 15,000-update endpoint are means across eight paired seeds. Effects use paired log ratios with Bonferroni-adjusted simultaneous 95% intervals. All eight pairs favor BHO at $N = 100$, and four of eight at $N = 800$.

N	Host	BHO	Change [simultaneous CI]	Holm p
100	0.5381	0.4754	-11.45% $[-17.15, -5.36]$	0.00197
800	0.0555	0.0569	$+2.19\%$ $[-10.19, 16.27]$	0.594

under the short schedule to 0.0082/0.0076 under the long schedule, while development errors rise from 0.4285/0.3896 to 0.5381/0.4754. Late development curves drift upward despite continued reductions in training error, indicating overfitting. During the continuation, host/BHO development errors rise by mean per-seed changes of $0.47\%/0.48\%$ at $N = 100$, with increases in all eight seeds for each arm, while training errors fall by $12.7\%/12.2\%$. At $N = 800$, development error falls further during the continuation, by mean per-seed changes of $1.47\%/1.44\%$ through update 18,000 (Figure 5).

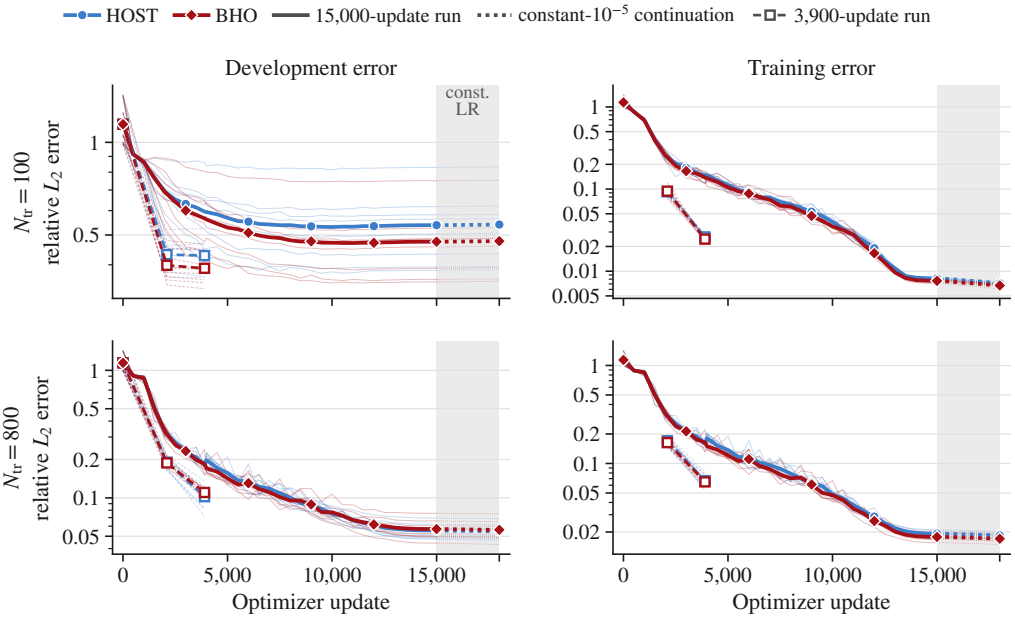


Figure 5: **The low-data advantage persists across two matched budgets.** Development and training relative L_2 errors for HOST and BHO; thick lines are means across eight paired seeds and thin lines show individual runs. The 3,900- and 15,000-update OneCycle schedules are separate trajectories from the same initial tensors. Short-schedule markers show only recorded checkpoints. The shaded region is the 3,000-update constant-10^{-5} continuation. At $N = 100$, late development error rises as training error falls; at $N = 800$, the continuation further lowers development error.

B.3 Shared-spectrum comparison

This comparison tests persistence of the shared-spectrum penalty with matched updates and its dependence on data or training duration. Mixed and homogeneous models use the same 3,900-update OneCycle horizon at $N = 100$ and 200, with seeds 1200–1207 and checkpoints at updates 2,100 and 3,900.

Each replicate samples an ordered 200-example subset without replacement; its first 100 examples form a nested core whose fitted normalizers serve all conditions. The mixed arm fixes the initial

signs, $D = \mathrm{diag}(s_0 \odot |\tanh(We)|)$; the homogeneous arm uses $|\tanh(We)|$. Per-head Fisher scalars are fixed before training from the first 16 core inputs and shared across sizes within each spectrum arm.

For each initialized layer and head, let a be the ordinary softmax probabilities over all keys and ρ the ungated residual, extended by zero on unscored keys. The fixed positive multiplier is

$$V = \mathbb{E}_{\text{calibration queries}} \mathrm{Var}_{j\sim a}(\rho_j), \qquad c = V^{-1/2}.$$

The fixed multiplier c scales the ungated residual before multiplication by the trainable gate, matching its initial mean categorical Fisher information. The multiplier retains the boundary-group offset and both self terms, but does not match complete-output tangents or later learning states.

Let $y_{N,t}$ be the mean paired log homogeneous/mixed contrast. The four scalar contrasts are persistence $P = y_{100,3,900}$, data moderation $D = y_{100,3,900} - y_{200,3,900}$, update moderation $U = y_{100,2,100} - y_{100,3,900}$, and interaction

$$I = (y_{100,2,100} - y_{200,2,100}) - (y_{100,3,900} - y_{200,3,900}).$$

Positive data and update contrasts indicate a smaller homogeneity penalty with more examples and more updates, respectively. The interaction I compares data moderation at 2,100 and 3,900 updates. Replicate-level intervals use 98.75% confidence, giving a Bonferroni-adjusted 95% family.

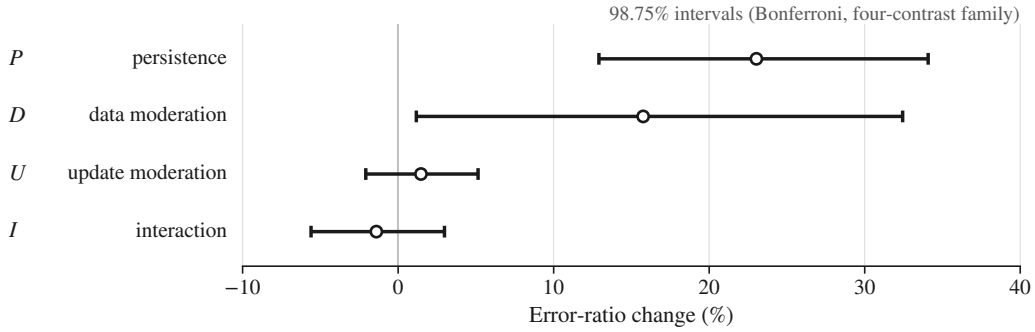


Figure 6: **Shared-spectrum comparison at a common update horizon.** Eight paired subset/initialization replicates. Error-ratio changes are $100[\exp(\bar{z}) - 1]$, where $\bar{z}$ is the mean paired log contrast; the zero line denotes a ratio of one. The homogeneity penalty persists at 3,900 updates; the lower confidence bound for the data-size ratio is below the prespecified threshold of 1.05, and the update and interaction intervals include zero.

Figure 6 summarizes the four contrasts. The shared penalty persists at 3,900 updates: $\exp(P) = 1.2304$ $[1.1291, 1.3408]$. Data moderation gives $\exp(D) = 1.1576$ $[1.0117, 1.3244]$; its point estimate exceeds the prespecified 1.05 attenuation threshold, but its lower bound does not. Update moderation and its interaction give $\exp(U) = 1.0147$ $[0.9792, 1.0514]$ and $\exp(I) = 0.9860$ $[0.9440, 1.0298]$. The update and interaction intervals include a ratio of one. At 3,900 updates, arithmetic mean mixed/homogeneous errors are 0.3311/0.4075 for $N = 100$ and 0.2365/0.2517 for $N = 200$; both models improve with more data.

B.4 Are learned coordinate signs necessary?

Fixing initial coordinate signs separates access to a mixed spectrum from learning its sign assignment. In a separate eight-seed shared-model study at BENO $N = 100$, fixed initial signs are non-inferior to free signs within the prespecified 3% margin: the one-sided upper bound on excess error is 0.43% (Table 12). Freezing magnitudes as well gives an upper bound of 3.85%, which does not establish non-inferiority.

Table 12: **Fixed-sign ablation in the shared model.** Separate BENO $N = 100$ block, eight paired seeds. The last column is the one-sided 95% upper bound on excess error relative to free signs; the non-inferiority margin is 3%. A missing bound denotes no such comparison.

Spectrum policy	Mean error	Non-inferiority bound
Free signs and magnitudes	0.3305	Reference
Fixed initial signs	0.3280	$+0.43\%$
Fixed initial spectrum	0.3371	$+3.85\%$
Homogeneous	0.4234	—

C ALGEBRA OF THE SCORE RESTRICTIONS

All statements concern free real hidden vectors and constant D. They are standard rank-factorization and inertia facts used to define the ablations, not a classification of the complete trained network on its reachable hidden states. Inertia counts positive, negative, and zero eigenvalues.

Proposition 1 (Separate and shared projections) *Fix diagonal $D \in \mathbb{R}^{r\times r}$ of rank m. Then*

$$\{Q^\top DK : Q, K \in \mathbb{R}^{r\times d}\} = \{A \in \mathbb{R}^{d\times d} : \operatorname{rank} A \le m\}.$$

If $P \in \mathbb{R}^{r\times d}$ has full row rank and D has inertia (p, q, z), then $P^\top DP$ has inertia $(p, q, z + d - r)$.

Proof. The rank upper bound follows from the product. Given $A = UV^\top$ of rank $k \le m$, place $U^\top$ in k nonzero rows of Q, put the corresponding rows of $V^\top$ divided by those entries of D in K, and set other rows to zero. For the shared case, choose invertible T with $PT = [I_r\ 0]$. Then $T^\top P^\top DPT = \operatorname{diag}(D, 0_{d-r})$, and congruence preserves inertia (Sylvester, 1852). Full row rank is essential; for example, $P = (1, 1)^\top$ and $D = \operatorname{diag}(1, -1)$ give zero. □

Full residual. For directed projections,

$$(Qx - Ky)^\top D(Qx - Ky) = x^\top Q^\top DQx - 2x^\top Q^\top DKy + y^\top K^\top DKy.$$

If $D \succeq 0$, its joint coefficient matrix $\lambda[Q, -K]^\top D[Q, -K]$ is positive semidefinite (PSD) or negative semidefinite (NSD) according to the sign of the scalar gate. Signed D can instead represent $x^2 - y^2$ with two latent coordinates. Homogeneity therefore restricts the full directed residual as well as the shared residual. Arbitrary isolated cross-couplings do not imply arbitrary coupled self terms. Ordinary dot-product attention does not remove this distinction, and a collection of heads with signed gates is not a globally PSD model.

Fixed signs. A symmetric M of rank $p + q$ and inertia $(p, q, d - p - q)$ has the spectral factorization $M = F^\top J_{p,q} F$, where $J_{p,q} = \operatorname{diag}(I_p, -I_q)$ and F has $p + q$ rows. This is minimal by the rank bound. A larger fixed signature can represent M if it has enough positive and negative slots; coordinate positions can be permuted with rows of F. Nonzero constant magnitudes can likewise be absorbed by row scaling (Kim et al., 2019). Such rescaling can change parameter norms, optimization trajectories, and finite-training error.

Attention support. With constant D, the query-self term is common to boundary keys. In B1's joint attention it shifts that group relative to grid keys and can alter boundary access. On the boundary-only routes in B2 and B3, it is common to the whole row and cancels. The key-self bias and cross term remain available for within-boundary selection. Under local D_j, the query-self term can also vary across keys. Thus the same residual can affect both group access and within-group selection under joint attention, whereas boundary-only softmax removes common group shifts.

D OLBC SCORE CONTROLS AND FINITE BOUNDARY RESPONSE

Population and interior support. The OLBC `poisson-circle-bc1` dataset is released at DOI `10.5281/zenodo.18376965` and contains 16,859 domain nodes and 1,004 Dirichlet

boundary nodes; Neumann and Robin index groups are empty. All 256 official test examples share the same geometry, coordinates, source, boundary indices, and Dirichlet type; only the prescribed boundary values vary. The same input structure holds for the 900 training-file examples used below.

Fields are interpolated to a 32×32 cell-center grid on $[-1, 1]^2$, and 128 boundary tokens are subsampled by polar angle. The released signed distance is positive inside this circle, so the inside input channel and training/evaluation mask use $\mathrm{sdf} > 0$, selecting 803 nodes with uniform norm weights for every arm and block. The fraction of grid nodes where this mask agrees with the unit-disc radius test is 0.9912.

Matched score interventions. The reduced-resolution B1-style joint-attention grid surrogate has 11 input channels, four width-128 blocks, eight heads, and 546,817 parameters. With the host's actual q_i, k_j, all added scores act on the same 128 boundary keys and all 1,152 query rows after ordinary score scaling:

$$r_{ij}^{\mathrm{GATE}} = a_{\ell,h}, \qquad r_{ij}^{\mathrm{BILINEAR}} = -2\lambda_{\ell,h} q_i^\top D_j k_j, \qquad r_{ij}^{\mathrm{BHO}} = \lambda_{\ell,h}(q_i - k_j)^\top D_j (q_i - k_j).$$

Each is zero on non-boundary keys. Both conditioned arms use $D_j = \mathrm{diag}(\tanh(W_{\ell,h} e_j))$ and $e_j = [n_{x,j}, n_{y,j}, \mathrm{bctype}_j, g_j]^\top$ from the same boundary-token rows as keys and values. BILINEAR removes only the two self terms; it retains BHO's local conditioner and 2,080 added parameters. GATE introduces 32 scalars. Under joint softmax, it modifies boundary-versus-grid attention mass while leaving conditional within-boundary weights unchanged at fixed host activations.

All four arms share the host initialization; zero-initialized a and λ preserve the same initial predictor and host gradients. BILINEAR and BHO start from the same W entries drawn from $\mathcal{N}(0, 1/4)$ (variance $1/4$) and train their conditioners separately.

Training, pairing, and evaluation role. In each of eight paired blocks, with seeds 53000–53007, 100 training rows are sampled without replacement from `train.nc[0:800]`; subsets may overlap across blocks. All four arms share the subset, normalizers, host initialization, and batch order within each block. Source, boundary-value, Robin-coefficient, and target scales are fitted on that block's 100 examples; the identically zero Robin coefficient uses scale one. Rows 800–899 are used to monitor training without model selection, and all 256 official `test.nc` examples are evaluated. This population and its official test split had been used in earlier exploratory analyses, so this study is not an independent validation on a previously unused population.

Each arm trains for 300 epochs and 2,100 AdamW updates with batch size 16, maximum learning rate 10^{-3}, weight decay 10^{-5}, a OneCycle schedule, and gradient clipping at 1.0. Training minimizes mean masked relative L_2 error in normalized units, with safeguard 10^{-8}. The terminal checkpoint is evaluated without selection by test or response error. Validation error decreased over the final recorded interval in every block and arm, with each arm's mean lowest at the terminal checkpoint.

Finite response and statistical families. Predictions are decoded to physical units before computing errors or differences. The primary endpoint averages untrimmed per-example relative L_2 error over 256 examples on the fixed 803-node mask. For the secondary endpoint, a NumPy default generator with seed 20260917 permutes the 256 row indices; consecutive entries form 128 disjoint pairs, reused across every arm and block. All pairs are retained, and pairing is independent of predictions and response magnitude. The field visualization uses the first pair in this ordering and the lowest block seed. For each pair p, define

$$\Delta\hat{u}_p = \hat{u}(g_{p,2}) - \hat{u}(g_{p,1}), \qquad \Delta u_p = u(g_{p,2}) - u(g_{p,1}).$$

The response endpoint pools squared norms before taking the relative error,

$$E_{\mathrm{response}} = \left(\frac{\sum_{p=1}^{128} \|\Delta\hat{u}_p - \Delta u_p\|_{2,\Omega}^2}{\sum_{p=1}^{128} \|\Delta u_p\|_{2,\Omega}^2} \right)^{1/2}, \tag{5}$$

with a 10^{-8} denominator safeguard. Fields are not independently centered, and masks and weights are identical for both pair members. All pairs have nonzero true-response norms; no pairs are excluded.

Examples or pairs are aggregated within each block before forming $\log(E_{\mathrm{BHO}}/E_X)$ for $X \in \{\mathrm{HOST}, \mathrm{GATE}, \mathrm{BILINEAR}\}$. The eight blocks are the inferential units, conditional on the common training pool and test set. Two-sided paired-t tests use seven degrees of freedom and Holm correction across the three prespecified primary solution contrasts; the three secondary response contrasts form a separate Holm family. Confidence intervals are pointwise 95% intervals on the mean log ratio, exponentiated to the ratio scale and converted to percentage changes. They are not simultaneous intervals (Table 13).

Table 13: **OLBC bc1 score-control contrasts.** Eight paired subset/initialization blocks. Percentage changes are calculated from mean paired log ratios; intervals are pointwise 95%, while Holm correction is applied separately within each three-contrast family. The last column counts blocks with lower BHO error. The primary BHO/GATE interval excludes zero, but its Holm-adjusted test does not reject the null hypothesis.

Contrast	Change [pointwise CI]	Raw p	Holm p	BHO lower
Primary solution error				
BHO/HOST	-6.86% $[-13.38, +0.15]$	0.0538	0.0635	6/8
BHO/GATE	-7.65% $[-13.93, -0.92]$	0.0318	0.0635	7/8
BHO/BILINEAR	-10.93% $[-15.91, -5.64]$	0.00209	0.00627	8/8
Secondary finite boundary-response error				
BHO/HOST	-9.68% $[-16.64, -2.14]$	0.0199	0.0348	7/8
BHO/GATE	-10.69% $[-18.08, -2.64]$	0.0174	0.0348	7/8
BHO/BILINEAR	-13.66% $[-19.30, -7.62]$	0.00134	0.00401	8/8

BHO reduces primary solution error relative to BILINEAR; the HOST and GATE contrasts establish neither superiority nor equivalence. All three response contrasts survive their separately specified correction, but solution and response errors use the same trained predictions. The response endpoint measures finite boundary-induced solution changes rather than boundary flux or PDE residuals.

Affine task-difficulty reference. As a post-hoc diagnostic, a multi-output affine ridge regression model is fitted using the same eight 100-example training subsets, 128 boundary-value observations, 32×32 target grid and 803-node mask as the neural comparisons. Regularization is selected by five-fold cross-validation using training examples only. Evaluation uses the same test set and 128 response pairs, with solution and response errors defined above. The small errors in Table 14 describe agreement with the stored targets on this fixed-source population. With geometry and source fixed, the linear Poisson solution depends affinely on the Dirichlet values. The diagnostic therefore characterizes this particular boundary-data distribution and is separate from the neural score arms and their Holm families.

Table 14: **Affine-ridge task-difficulty diagnostic on OLBC bc1.** Same eight training-subset blocks; mean test relative-L_2 error and pooled finite-response error. The solution interval is a pointwise 95% t interval; this post-hoc reference is separate from the neural comparisons and their multiplicity adjustments.

Reference	Test solution error	Solution 95% CI	Pooled response error
Affine ridge	3.43×10^{-8}	$[3.40, 3.47] \times 10^{-8}$	3.69×10^{-8}

D.1 Boundary-response pairing sensitivity

Sensitivity to the prespecified pairing is assessed using all unordered pairs of the same $M = 256$ test examples. For prediction errors $e_i = \hat{u}_i - u_i$, let $\bar{e}$ and $\bar{u}$ denote means across examples at each node, not spatial means. The weighted norm uses uniform weights on the same 803-node interior

mask. The all-pairs identity is

$$E_{\mathrm{all}}^2 = \frac{\sum_{i<j} \|e_i - e_j\|_w^2}{\sum_{i<j} \|u_i - u_j\|_w^2} = \frac{\sum_i \|e_i - \bar{e}\|_w^2}{\sum_i \|u_i - \bar{u}\|_w^2}.$$

The reported metric is the positive square root E_{all}. Its denominator is finite and positive, with no added epsilon; the original 128-pair safeguard is unchanged. These means define the metric and do not alter predictions. This is a post-hoc sensitivity analysis of the same decoded predictions, not a new replication. The eight paired training blocks remain the inferential units; the 32,640 pairs are not independent replicates.

Table 15 compares the original response family with a separate three-contrast Holm sensitivity family. All three all-pairs contrasts survive that adjustment, and their effect estimates differ from the original by less than 0.5 percentage points. On this population, the relative improvements in all-pairs response closely track reductions in pooled solution error, so this analysis is not interpreted as separate mechanistic evidence.

Table 15: **Boundary-response sensitivity to pairing.** Eight paired blocks; changes use paired log ratios with pointwise 95% confidence intervals. The 128-pair secondary response contrasts and the all-pairs sensitivity contrasts use separate three-contrast Holm families.

	128 fixed pairs		All pairs	
Contrast	Change [pointwise CI]	Holm p	Change [pointwise CI]	Holm p
BHO/HOST	-9.68% $[-16.64, -2.14]$	0.0348	-9.21% $[-16.11, -1.73]$	0.0386
BHO/GATE	-10.69% $[-18.08, -2.64]$	0.0348	-10.39% $[-17.75, -2.36]$	0.0386
BHO/BILINEAR	-13.66% $[-19.30, -7.62]$	0.00401	-13.49% $[-18.99, -7.62]$	0.00367

D.2 Native Boomerang score controls

The native experiments evaluate the released XGAOT with learned boundary extension at commit `82094ffe2d98` on `poisson-boomerang-bc4` (Mixed, DOI `10.5281/zenodo.18377507`) and `poisson-boomerang-bc5` (Mixed+, DOI `10.5281/zenodo.18377531`). Both have fixed non-convex geometry, 15,187 valid native nodes, and 1,160 boundary nodes, with Dirichlet, Neumann, and Robin data. Mixed varies the boundary conditions while keeping the source fixed; Mixed+ varies both. Models are trained separately on each population.

The four arms use the same boundary-aware host. GATE multiplies the boundary message by $\exp(a)$; it is not the uniform logit bias that cancels under boundary-only softmax. BILINEAR adds a locally conditioned cross-only residual; BHO adds the full directed quadratic. Their head gates are zero-initialized and freely signed, and they use the same conditioner architecture with parameters trained separately. Host initialization, preprocessing statistics, batch streams and checkpoint-selection procedures are matched across arms within a block.

The conditioner inputs are $e_j = (n_{x,j}, n_{y,j}, t_j, \tilde{g}_j)^\top$: outward-normal components, a type code $t_j = +1, -1, 0$ for Dirichlet, Neumann, and Robin nodes, and the metadata-defined boundary datum, standardized using that segment's block-training statistics. The Robin coefficient is excluded from this conditioner. All four arms receive this same four-vector through the host input path, appended to the released eight-channel boundary token (coordinates, signed distance and its gradient, and the normalized boundary-condition tuple). Normals are $-\nabla\mathrm{sdf}/\|\nabla\mathrm{sdf}\|$. The boundary route has six layers, four heads, and head dimension 128.

Robin preprocessing note. The released host reverses the metadata-defined Robin datum and coefficient before coefficient-aware normalization, whereas the score conditioners use the metadata-

defined datum. The comparisons therefore concern this encoding. Correcting it could change the relative performance of the arms; that comparison has not been evaluated.

Training and evaluation. Each population uses eight paired subset/initialization blocks with $N_{\mathrm{tr}} = 1,024$, with seeds 62000–62007 for Mixed and 63000–63007 for Mixed+. Subsets contain distinct rows drawn from training rows $[0, 8{,}192)$; subsets may overlap across blocks. Validation uses rows $[8{,}192, 8{,}448)$, and evaluation uses all 256 official test examples. Every arm receives 24,000 updates with batch size 8, corresponding to 192,000 example presentations, with adaptive gradient clipping at ratio 0.5. Preprocessing statistics are fitted on each block's 1,024 training examples. Training uses relative L_2 loss on centered, normalized outputs with the released 0.2% node trimming, whereas the primary test metric is computed from decoded outputs without trimming. The earliest checkpoint attaining the lowest validation error under the released metric, which trims errors and takes their median over examples, is selected. The primary test results use these selected weights, which need not be the terminal weights. Test evaluation follows checkpoint selection.

The primary test endpoint is the mean of per-example untrimmed relative L_2 errors in decoded physical units, with denominator safeguard 10^{-8}. Norms use uniform weights over all 15,187 valid native nodes, including boundary nodes, without interpolation or the circle experiment's interior mask. Table 16 gives arithmetic means and SDs of the eight block endpoints. The effects in Table 5 are instead calculated from mean paired log ratios; they are not ratios of these arithmetic means. Each population has its own three-contrast Holm family, conditional on its training pool and common test set. Blocks, rather than nodes, examples, or checkpoints, are the inferential units, and the displayed intervals are pointwise 95%.

Table 16: **Absolute errors on the native Boomerang populations.** Arithmetic mean ± SD across eight paired blocks, using validation-selected weights and the primary untrimmed mean test metric.

Arm	Mixed	Mixed+
HOST	0.1691 ± 0.0072	0.2256 ± 0.0078
GATE	0.1680 ± 0.0056	0.2235 ± 0.0085
BILINEAR	0.1711 ± 0.0055	0.2258 ± 0.0079
BHO	0.1662 ± 0.0068	0.2269 ± 0.0073

In a descriptive sensitivity analysis, the secondary released trimmed/median test metric gives BHO/HOST, BHO/GATE, and BHO/BILINEAR changes of -2.17%, -1.37%, and -2.16% on Mixed and -2.96%, -1.61%, and -1.71% on Mixed+. These summaries differ in trimming and aggregation from the primary untrimmed mean endpoint and do not replace its conclusions.

E Training and implementation details

Host routes. B1 uses joint self-attention. In B2, the grid stream self-attends and receives boundary information through per-block cross-attention. B3 uses 32 learned slices, slice attention, and broadcast, with a direct boundary cross-attention route. B2 and B3 are reduced GNOT-style and Transolver-style hosts, respectively.

B2 is trained for 300 epochs with batch size 16, corresponding to 15,000 updates at $N_{\mathrm{tr}} = 800$, with AdamW/OneCycle, pool-fitted preprocessing, and terminal development evaluation without checkpoint selection. Across eight paired runs, HOST and BHO obtain mean errors 0.2342 and 0.1236, respectively, corresponding to a 47.2% reduction by the ratio of arithmetic means; BHO error is lower in all eight pairs. The exact Wilcoxon test gives $W = 0$, $p = 0.0078$ (Table 7). A post-hoc paired-run bootstrap with 100,000 resamples gives a 95% interval of $[41.9, 52.2]\%$ for the same estimand.

BENO boundary inputs. Each example supplies 128 boundary tokens, used in their released order without resampling. At each token, a tangent is formed from its two nearest boundary neighbors; a perpendicular unit vector is oriented away from $(0.5, 0.5)$ to estimate the normal. Boundary tokens carry coordinates, a boundary indicator, the scaled datum, the estimated normal, and the type code;

grid tokens carry coordinates, the scaled source, and interior/exterior indicators. Unused slots are zero. The type code is constant at -1 for the Neumann-labeled population. Source, boundary, and target scaling is divide-only, using standard deviations over all grid nodes, boundary tokens, and physical-interior target nodes, respectively. The score conditioner uses the same scaled boundary datum as the host.

Optimization and precision. The primary host has a multilayer perceptron (MLP) expansion ratio of 2 and 546,689 parameters without BHO. The BENO comparisons use AdamW with maximum learning rate 10^{-3}, weight decay 10^{-5}, batch size 16, and the specified OneCycle schedule. The primary and longer matched-budget comparisons and the shared-spectrum equal-update control use $\mathrm{pct_start} = 0.3$, cosine annealing, β_1 cycling between 0.85 and 0.95, division factor 25, and final division factor 10^4. They use 32-bit floating-point (FP32) tensors with TensorFloat-32 (TF32) enabled and no automatic mixed precision (AMP). Both matched-budget comparisons use $\beta_2 = 0.999$, AdamW $\epsilon = 10^{-8}$, and no gradient clipping or dropout. Initial tensors and batch sequences are matched within pairs; GPU arithmetic may remain nondeterministic. The bc1 score-control models use FP32 tensors with TF32 disabled.

Initial predictor and score scale. The full-versus-cross-only comparisons in BENO, OLBC bc1, native Boomerang, and Steel–Circle^H use no residual-scale calibration. Zero gates match the initial predictors and make the added-score derivatives with respect to conditioner parameters vanish. The ungated residuals and initial gate derivatives nevertheless differ. These comparisons therefore change the learning parameterization without isolating self terms from score scale or optimization. The categorical Fisher calibration in Appendix B.3 applies only to the separate shared-spectrum comparison.

Paired endpoints and figure data. The accompanying files contain paired endpoints and contrast families, the all-pairs response analysis, native Boomerang block-level results, and inputs to the tables and figures. The response visualization uses physical-unit fields and block-level pooled errors; learning curves use checkpoint errors; the paired BENO score-control figure uses the eight paired endpoints and the pointwise and family-simultaneous intervals.

E.1 Computational cost

The primary BENO host and locally conditioned, all-query BHO are profiled on one NVIDIA A100-SXM4-80GB, using trained $N_{\mathrm{tr}} = 100$ checkpoints from the 3,900-update comparison and the same device-resident training batch. The batch contains 16 examples, each with 1,024 grid and 128 boundary tokens and 10 input features. Both arms use PyTorch eager execution in FP32 without AMP, with TF32 enabled in the backend settings and explicit attention rather than fused attention. The BHO gates are nonzero, so the added score path is active.

Table 17: **Cost of the primary BENO score intervention.** Batch size 16 on one A100; latency is the median of five repetition means. Memory is peak allocated device memory, including the resident model and batch, and is identical across the five repetitions for each arm and mode. Ratios are computed before rounding.

Measurement	HOST	BHO	BHO/HOST
Inference (ms/batch)	15.846	24.669	1.56×
Forward–loss–backward (ms/batch)	39.613	63.164	1.59×
Inference peak allocated (MiB)	1,416.0	1,497.2	1.06×
Forward–loss–backward peak (MiB)	4,952.5	5,437.6	1.10×

Table 17 reports medians of five repetition means, each computed from 100 synchronized CUDA-event timings after 20 warm-up iterations. Each repetition loads one model in a fresh process. At the start of each repetition, no other GPU compute processes were recorded. Inference latency covers evaluation-mode forward computation without gradients. Training computation includes the forward pass, interior relative-L_2 loss, and backward pass, including gradient-buffer allocation. Gradient

clearing, optimizer and scheduler steps, loading, preprocessing, and device transfer are outside the timed region. Peak allocated memory includes the resident model and batch; reserved allocator memory is excluded. These measurements concern this implementation and configuration, not end-to-end training time. The 0.38% parameter increase therefore does not imply a comparable increase in runtime or memory.

F STEEL–CIRCLE^H ELASTICITY SCOPE COMPARISON

The four-arm Steel–Circle^H comparison tests the score intervention on linear elasticity. It uses the released `elasticity-circlehollow-m1` task and a common XGAOT host with the same eight added boundary features in every arm; GATE scales the boundary attention message, while BILINEAR and BHO add their conditioned cross-only and full quadratic scores. Each of eight paired subset/initialization blocks uses 1,024 training examples and 12,000 updates, with terminal-weight evaluation on 256 test examples. The primary endpoint is the mean per-example joint-displacement relative L_2 error, computed from decoded predictions without trimming over all finite-target nodes, including boundary nodes.

The added feature vector is

$$e_j = [n_x, n_y, b_x, b_y, m_{Dx}\tilde{u}_x, m_{Dy}\tilde{u}_y, m_{Nx}\tilde{\tau}_x, m_{Ny}\tilde{\tau}_y]^\top,$$

where b_x, b_y encode prescribed displacement $(+1)$, traction (-1), or neither (0); m_D and m_N select the prescribed displacement and traction components, respectively. Each displacement and traction component is standardized with that component's block-training statistics. Normals are obtained from a local PCA tangent and oriented using $-\nabla\mathrm{sdf}$. The host uses two component-wise boundary-extension passes with shared parameters; each receives the same eight added features. Both score conditioners use this eight-vector and head dimension 128. Each layer and head has a 128×8 conditioning matrix, with entries initialized from $\mathcal{N}(0, 1/8)$ (variance $1/8$).

Table 18 reports the four primary means and three primary contrasts. BHO has higher joint-displacement error than HOST and GATE; its contrast with BILINEAR is unresolved. This result limits the intervention's HOST-relative utility in the tested elasticity setting.

Table 18: **Steel–Circle^H: joint-displacement error.** Panel A gives arithmetic means ± SD across eight blocks. Panel B gives paired-log percentage changes with pointwise 95% t intervals and Holm-adjusted p-values for the three-contrast primary family; positive changes mean higher BHO error.

A. Primary joint-displacement error

Arm	Mean ± SD
HOST	0.3565 ± 0.0062
GATE	0.3552 ± 0.0073
BILINEAR	0.3659 ± 0.0069
BHO	0.3675 ± 0.0079

B. Primary comparison family

Contrast	Change [pointwise 95% CI]	BHO lower	Holm p
BHO/HOST	$+3.06\%$ $[+1.65, +4.49]$	0/8	0.00257
BHO/GATE	$+3.46\%$ $[+2.09, +4.85]$	0/8	0.00158
BHO/BILINEAR	$+0.43\%$ $[-1.15, +2.04]$	4/8	0.539