EXECUTABLE INTERVENTION GAMES: FROM ATTRIBUTION TO PHYSICAL CONSEQUENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Explainable AI helps researchers formulate scientific hypotheses from predictive models, but fidelity to a model does not establish that the relationships it has learned hold in the physical system. Testing these relationships is difficult when input perturbations, such as masking, do not correspond to the physical actions of interest. To bridge this gap, we introduce executable intervention games, in which a model and a numerical reference answer the same questions about prescribed actions executed from shared starting conditions. Comparing their conditional action effects and Shapley attributions reveals response errors that inform researchers' choices of action coverage or response supervision. In heat conduction, supervising single-action temperature differences improves held-out effect and attribution accuracy over longer field training and absolute-temperature supervision with identical labels. It also reduces average error in predicting how other active actions modify an effect in new materials, even though this dependence is not directly supervised. In cylinder flow, covering action combinations improves responses beyond controls with matched training states and histories, including at held-out Reynolds numbers. By preserving the physical question through explanation, feedback, and evaluation, executable games make learned relationships testable and support their improvement for scientific inquiry.

1 INTRODUCTION

Machine-learning models can suggest relationships for scientific investigation, and explainable AI (XAI) helps researchers inspect what those models have learned. Attribution maps, for example, have been used to identify influential structures in fluid flow (Cremades et al., 2024). Such explanations often motivate a physical question: would changing a forcing, a boundary or a material property affect the outcome as the model suggests? Model fidelity differs from fidelity to the data-generating process (Sun et al., 2024). For science, this means a faithful explanation may describe an incorrect physical relationship.

To test that relationship, the explanation and reference experiment must address the same change. Consider flow around a cylinder, where the fluid exerts a resisting force called drag. Masking a vortex, a region of rotating fluid, edits the observation supplied to a predictor. Rotating the cylinder wall instead changes the physical system and allows the flow to respond. Figure 1 illustrates why changing an observation and changing the system define different experimental questions.

We introduce *executable intervention games* to connect explanations to physical checks and learning feedback. A game fixes the starting conditions, action programs and measurement, so the model and a numerical reference answer the same action questions. We compare the effect of adding an action with different other actions already present, then use the same Shapley rule to summarize these effects as attributions. Building on removal-based explanations and interventionally consistent surrogates (Covert et al., 2021; Dyer et al., 2024), this shared experiment lets a researcher locate an incorrect response relationship, choose training feedback and test whether that relationship improves.

Heat conduction provides a complete instance. A network predicts temperature throughout a material, and local conductivity changes alter how heat passes through it. We check the resulting temperature differences against a partial differential equation (PDE) reference and use the discrepancies to define

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.

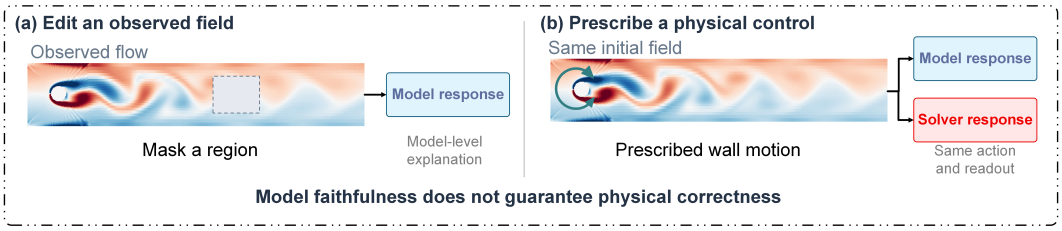


Figure 1: **Explaining a model and checking its physics ask different questions.** (a) Masking an observed field probes the model response to an edited input. (b) Prescribed wall motion defines a shared model–solver experiment with the same action and readout. Both panels show the same Re80 pre-control vorticity field. The mask and wall-motion arrow are schematic, and the solver provides a numerical physical reference.

response supervision. Ordinary field training and absolute-temperature supervision receive identical endpoint fields, allowing us to test what the model learns from the same labels.

The conduction study also reveals an improvement beyond the directly supervised relationship. Although response supervision constrains only single-action effects against a no-action background, it reduces the average error in predicting how other active actions modify those effects in new materials. Matched labels and independent weight selection test the effect of supervision. In cylinder flow, covering on/off action combinations improves response prediction after matching training flow states and histories, with gains shared by several coverage menus.

These studies connect explanation to researcher-selected feedback and controlled tests of held-out physical responses. Complementary studies distinguish field completion from executed actions, show how exact attribution can summarize an incorrect learned response, and test feedback in smoke. Executable predictions also supply hypotheses for separate numerical checks.

2 Executable Intervention Games

To check a predicted relationship, we first need an action that both the model and reference can evaluate. Consider a material held hot on its left edge and cold on its right. Increasing conductivity locally changes the equilibrium temperature field. A field predictor and a PDE solver evaluate the same modified material, and averaging each field over the same region of interest (ROI) gives comparable measurements (Figure 2).

2.1 Defining the shared experiment

A *parent* p specifies the configuration before intervention and any relevant history: the original conductivity and boundaries in conduction, or the starting state and history in a flow. We fix m admissible action programs b_i, their amplitudes α_i and combination rule, the prediction horizon and a scalar measurement Y. For active programs $S \subseteq N = \{1, \dots, m\}$, let $G_{p,\theta}(S)$ and $G_{p,\mathrm{ref}}(S)$ denote the model and reference measurements. Subtracting each one's no-action outcome defines the parallel games

$$v_{p,k}(S) = G_{p,k}(S) - G_{p,k}(\emptyset), \qquad k \in \{\theta, \mathrm{ref}\}. \tag{1}$$

Each action retains its amplitude as others are added, and every combination must remain admissible. A player is therefore a prescribed physical program, and the game value is the measured response of a combination.

In conduction, four local increments to log-conductivity define sixteen configurations. The model predicts each 32×32 temperature field and Y averages its ROI. Other predictors can directly output a force statistic or forecast a future field. The required match concerns the action, history and readout, while the input representation can vary (Appendix B.1).

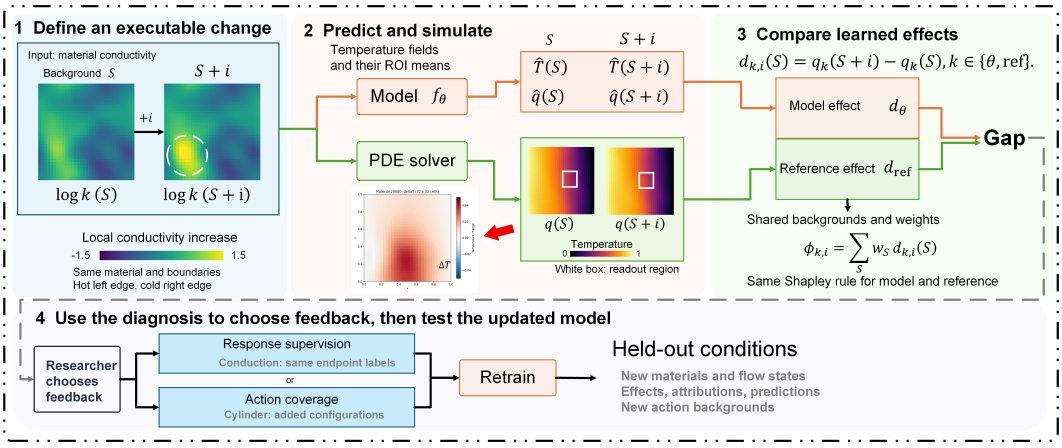


Figure 2: **Shared response checks guide training feedback and held-out evaluation.** (1–2) Model and reference evaluate the same conductivity change and ROI-temperature readout. The 32×32 reference fields come from a pilot material, while model outputs are symbolic. The dashed circle marks one Gaussian standard deviation. (3) Endpoint differences give $d_{k,i}(S)$, and shared backgrounds and Shapley weights define attributions. Here k denotes model or reference, $q_\theta = \hat{q}$ and $S + i = S \cup \{i\}$. (4) The researcher chooses action coverage or response supervision, which reuses endpoint labels. Held-out conditions test effects, attributions and predictions.

2.2 Checking effects and explaining their allocation

A local conductivity change can have a different effect when other regions have already changed. We call those other active actions the *background*. Adding action i to background S has conditional effect

$$d_{p,i}(S) = v_p(S+i) - v_p(S), \qquad i \notin S, \quad S + i = S \cup \{i\}. \tag{2}$$

Both games use this operation, so we suppress their index. Comparing $d_{p,\theta,i}(S)$ with $d_{p,\mathrm{ref},i}(S)$ checks the learned relationship in that background. Its change when action j is added gives the interaction $I_{p,ij}(S) = d_{p,i}(S+j) - d_{p,i}(S)$ for distinct $i, j \notin S$.

To allocate the total response among actions, we use the standard Shapley rule. For a fixed parent, it averages each action's effect over all orders in which actions can be added:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(m - |S| - 1)!}{m!}\, d_i(S). \tag{3}$$

With two actions, the first receives the average of its effects without and with the second. The game fixes the experimental question, conditional effects check individual responses, and attribution summarizes them across backgrounds.

Using the same rule for model and reference separates explanation approximation from learned-response disagreement. If $\widehat{\phi}$ approximates the exact allocation ϕ, then

$$\widehat{\phi}(v_\theta) - \phi(v_{\mathrm{ref}}) = \underbrace{\widehat{\phi}(v_\theta) - \phi(v_\theta)}_{\text{explanation approximation}} + \underbrace{\phi(v_\theta) - \phi(v_{\mathrm{ref}})}_{\text{model–reference disagreement}}. \tag{4}$$

Exact enumeration removes the first term. Averaging can still cancel conditional-effect errors, and reducing those errors need not reduce attribution error. We therefore check responses, conditional effects and attributions together (Appendix B.3).

A selected effect needs two endpoint measurements, which can be reused across checks. Full enumeration evaluates 2^m configurations and supplies all effects and exact Shapley values. We enumerate our four-, six- and eight-action games to study response disagreement without attribution-approximation error.

3 LEARNING FROM RESPONSE DISCREPANCIES

A diagnosed discrepancy gives the researcher a concrete learning target. Feedback can change which action outcomes training covers or how available outcomes supervise the model. We isolate these choices in conduction and cylinder flow, then test their consequences under held-out conditions.

3.1 SUPERVISING RELATIONSHIPS WITH THE SAME LABELS

A pair of conduction fields supplies both absolute ROI temperatures and their difference. Write q_0, q_i for the no-action and single-action temperatures, with hats for predictions. All arms share a field-prediction loss on paired outcomes and replayed no-action fields, $\mathcal{L}_{\text{field}}$. Paired-field training P uses this loss alone. Response supervision T emphasizes the difference, while absolute-temperature supervision A emphasizes the endpoints:

$$\begin{aligned} \mathcal{L}_P &= \mathcal{L}_{\text{field}}, \\ \mathcal{L}_T &= \mathcal{L}_{\text{field}} + \lambda_T \frac{[(\hat{q}_i - \hat{q}_0) - (q_i - q_0)]^2}{\sigma_B^2}, \\ \mathcal{L}_A &= \mathcal{L}_{\text{field}} + \lambda_A \frac{(\hat{q}_0 - q_0)^2 + (\hat{q}_i - q_i)^2}{2\sigma_B^2}. \end{aligned} \tag{5}$$

Here σ_B^2 is the training-temperature variance and losses are averaged over pairs. Within each fit, the arms share initialization, fields, normalization and sampling. T reuses endpoint labels without additional reference solves. We set $\lambda_T = 1$ and calibrate λ_A to match the initial auxiliary-gradient norm in the long-budget comparison (Appendices B.4 and E).

The difference between these objectives is easiest to see in the endpoint errors $e_0 = \hat{q}_0 - q_0$ and $e_i = \hat{q}_i - q_i$. Apart from their common normalization, T penalizes $(e_i - e_0)^2$, whereas A penalizes

$$\frac{e_0^2 + e_i^2}{2} = \frac{(e_i + e_0)^2 + (e_i - e_0)^2}{4}. \tag{6}$$

A therefore allocates supervision to both a shared temperature offset and the response error. T concentrates its auxiliary term on the response, while the shared field loss still constrains absolute predictions. The labels contain both kinds of information; the objective changes their relative emphasis.

A separate rerun trains on new materials and gives T and A the same finite weight-selection procedure. A development panel selects one of three weights per arm before the selected fits are tested on new confirmation materials (Appendix E.3).

For each of 32 training materials, the no-action field and four single-action fields supply four singleton–null contrasts, giving 128 directly supervised effects. None compares an action against a nonempty background. Testing those backgrounds in new materials asks what the supervision teaches beyond the relations it explicitly penalizes.

3.2 COVERING THE ACTION COMBINATIONS BEING EXPLAINED

When training lacks the needed outcomes, feedback concerns action coverage. In cylinder flow, the initial data include the all-controls configuration, yet individual control effects remain inaccurate. We add partially active configurations through full-background leave-one-out, balanced size-four or random Boolean menus. Matched labels, updates and fitted seeds isolate the coverage comparison, with a parent/history-matched control separating action coverage from training-state support.

The researcher selects a menu using the diagnosed relationship and available labels. The cylinder full-background menu originated in earlier context-model diagnostics, and Appendix D.3 gives a retrospective CNN illustration. The comparisons test the resulting coverage choices.

3.3 EVALUATING WHAT THE FEEDBACK TEACHES

Held-out evaluation retains the action definitions and readout, measuring response, conditional-effect and attribution errors alongside field accuracy. In conduction, *background modulation*, $d_i(S) - d_i(\emptyset)$

Table 1: Experimental interfaces and units. ROI denotes a region of interest. Fitted seeds and multiple action combinations are repeated measurements within a physical case, not additional cases. Detailed splits are in Appendix B.

System	Predictor and readout	Actions and feedback	Held-out conditions and unit
Heat conduction	Conductivity and coordinates $\rightarrow$ temperature field; ROI mean	Four local conductivity changes; same labels, different supervision	New materials and nonempty action backgrounds; whole material
Cylinder flow	Pre-control history and coefficients $\rightarrow$ four force/vorticity readouts	Six wall-rotation programs; action coverage	Four fixed flow bases, three perturbation draws each; distinct pulse histories and Re strata
1D dynamics	State/history $\rightarrow$ future field; local energy	Forcing programs; query construction and coverage	Parent states; diagnosis and hypothesis panels kept separate
Smoke transport	Executed-prefix history $\rightarrow$ future field; downstream ROI mass	Eight force programs; coverage and supervision	New trajectories; whole trajectory

for nonempty S, isolates how an effect changes when other actions become active. We subtract each model's own singleton prediction before comparison with the corresponding reference quantity, directly testing a relationship absent from the supervision loss.

Fixing the parent, action, background, amplitude and measurement window also turns a predicted direction or interaction into a physical hypothesis. Separate reference experiments and numerical-sensitivity checks test these downstream predictions.

4 Experiments

We ask whether checking executable responses can guide feedback that improves held-out relationships. Conduction provides a same-label supervision test, and cylinder flow tests action coverage after matching training-state support. One-dimensional dynamics and smoke transport examine the explanatory target, feedback limits and physical hypothesis testing (Table 1).

We enumerate each small game and measure errors in null-relative responses, conditional effects and exact Shapley allocations. Conduction uses dimensionless temperature units. Normalized root-mean-square errors (nRMSE) elsewhere use study-specific reference scales. Appendix B details the paired comparisons and confidence intervals. Smoke uses PDEArena/PhiFlow and cylinder flow uses FluidGym with PICT (Gupta & Brandstetter, 2023; Holl et al., 2020; Becktepe et al., 2026; Franz et al., 2026).

4.1 Response checks guide same-label feedback in conduction

Each conduction material is evaluated under all sixteen combinations of four conductivity changes, with a U-Net predicting the equilibrium temperature field. On eight development materials, the initial seed-17 model has mean relative field error of 1.56%, while conditional ROI-effect RMSE averages 22.8% of each material's reference-effect RMS. The effect check identifies inaccurate response amplitudes despite close field predictions, giving us a specific target for supervision (Appendix E).

Training pairs show how changing one conductivity patch affects ROI temperature while the other patches remain unchanged. We also ask whether the model predicts that effect when other patches have already changed. The PDE is linear in temperature for a fixed material, but conductivity changes the operator, so combined effects need not add.

We compare field training P, absolute-temperature supervision A and response supervision T using identical labels and three paired fits (Eq. (5)). T and initially gradient-calibrated A receive 1,500 updates. P receives 1,590 updates and 6.2% more measured training time than T in this comparison.

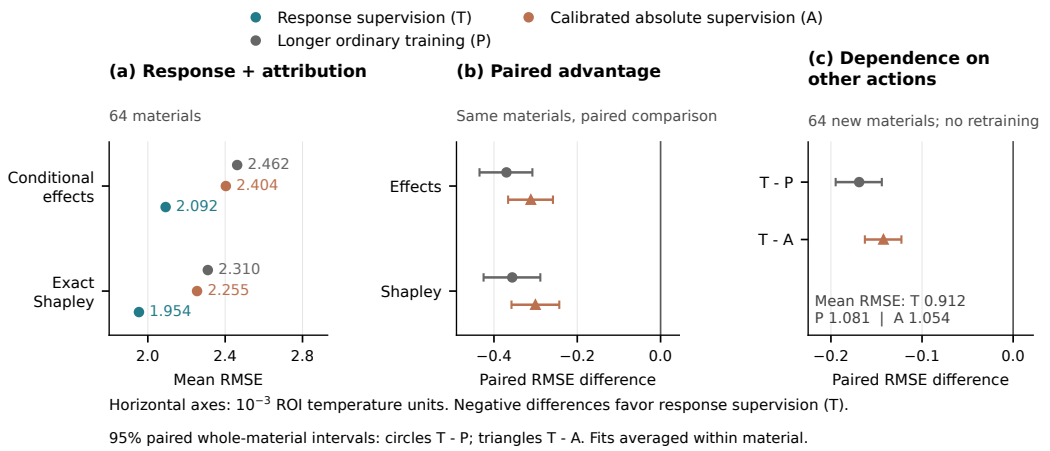


Figure 3: **Single-action supervision improves responses and their dependence on other actions.** (a–b) Effect/Shapley errors and paired differences on 64 long-budget test materials. Shapley is a secondary analysis. (c) Modulation error on 64 disjoint materials generated after fixing the models and analysis. Modulation measures 28 nonempty-background contrasts after subtracting each model's own singleton effect, although training supervises only singleton–null differences. P/A/T share labels and use the objectives in Eq. (5). Errors average three fits within material, with 95% whole-material intervals conditional on those fits. Axes use 10^{-3} temperature units, and negative differences favor T.

On 64 held-out materials, T reduces effect RMSE to .002092 from .002462 for P and .002404 for A, with both paired 95% intervals excluding zero. Exact-Shapley errors also fall (Figure 3a,b). Emphasizing the difference between endpoints therefore improves the learned relationships beyond longer field training and initially gradient-calibrated endpoint supervision. Mean controlled-field relative L_2 error also falls, to 1.048% from 1.071% for P and 1.098% for A: the response gain accompanies better average field prediction (Appendix E.2).

To test background dependence beyond direct supervision, we fix the models and modulation analysis, then generate 64 disjoint materials without retraining. Across the 28 nonempty-background contrasts $d_i(S) - d_i(\emptyset)$, T's modulation RMSE is .000912, compared with .001081 for P and .001054 for A (Figure 3c). Both paired 95% intervals exclude zero, and all three fit means favor T. Averaging fits within material, 60 of 64 materials favor T over P and 63 of 64 favor T over A.

Because modulation subtracts each model's own singleton effect, this gain concerns the predicted change with background. Each arm also outperforms replacing all its nonempty-background effects with its singleton prediction (Appendix E.4). All arms use background dependence, but T predicts it more accurately on average. Single-action supervision has thus improved a relationship it never directly penalized, in new materials.

An independent rerun tests whether the advantage persists when both auxiliary objectives receive the same weight search. With new training materials and baselines, T and A each select $\alpha = 4$ from $\{.25, 1, 4\}$ on sixteen development materials. On 64 new confirmation materials, T's effect RMSE is 19.4% lower than A's, with a paired difference of $-.000438$ (95% interval $[-.000498, -.000380]$). All three fit means and all 64 material means favor T. T also reduces mean modulation error relative to A and the longer-update P control (Table 2), with material and fit heterogeneity detailed in Appendix E.3. This independent comparison supports the response advantage after finite weight selection.

4.2 Action coverage improves responses beyond matched training support

In cylinder flow, six wall-rotation programs act on an evolving fluid, and a CNN predicts changes in mean drag, mean lift, endpoint lift and a near-wake vorticity-intensity proxy. We combine their errors using fixed training scales. Each arm adds 24 labels to 172 common outcomes and trains for 1,200 updates at three paired fitted seeds. Training uses Reynolds numbers 60, 80, 120 and 160, a parameter

Table 2: Independent conduction confirmation on 64 new materials. RMSEs use 10^{-3} temperature units, averaging three fits within material. Bold/underline mark lowest/second-lowest error means. T/A select $\alpha = 4$; times are per-fit training means (Appendix E.3).

Arm	Effect ↓	Modulation ↓	Time (s)
T	**1.816**	**.878**	20.63
A	<u>2.253</u>	.974	20.68
P	2.502	<u>.964</u>	18.38

Table 3: Cylinder coverage menus. Arms share 172 labels and add 24 outcomes, with 1,200 updates. Rφ matches their allocation across the training parents and histories in H/S/B. Boolean actions are off or at amplitude .08.

Code	Added action outcomes
R	Signed continuous coefficients on the original parents and histories
Rφ	Same coefficients on the parents and histories used by H/S/B
H	Full-background leave-one-out
S	Balanced four-active-action configurations
B1–B3	Three fixed random Boolean menus (B mean averages their errors)

Table 4: A learned control effect before and after coverage feedback. Entries are action-0 mean-drag effects in 10^{-3} units, not attributions. The Re80 example is selected retrospectively as the upper-median draw by H−R all-edge error; positive history and fit seed 17 are fixed (Appendix D.3). Both H effects improve over the initial CNN, while Rφ is closer than H on the three-control background. Figure 4 evaluates the overall coverage comparison.

Background	Reference	Initial CNN	R	Rφ	H
Three controls ($14 \to 15$)	.214	1.536	.974	.380	.419
Five controls ($62 \to 63$)	.331	2.762	1.745	.789	.123

comparing inertia with viscosity. Evaluation uses four fixed flow bases at Re60/80/100/170, with three perturbation draws per base and two pulse histories per draw. Re100 tests interpolation and Re170 extrapolation.

A retrospective initial-CNN example makes the coverage question concrete: its full-action drag response is close to the reference, yet its predicted individual control effect is 8.34 times the reference effect. Table 4 shows what changes after feedback. Relative to the initial model, full-background coverage improves this effect and an internal-background effect absent from its added supervision. The same action still increases drag, but the predicted magnitude becomes more realistic. This example illustrates why checking only the total response can miss a scientifically relevant error.

The controls in Table 3 distinguish seeing relevant flow states from seeing relevant action combinations. Executing the same continuous coefficients from matched parents and histories reduces effect nRMSE from .2336 to .1599. Covering on/off combinations improves further, to .1157 for full-background leave-one-out, .1180 for balanced size-four and .1145 for the mean of three random Boolean menus. Random Boolean coverage also lowers Shapley error to .08451 from the matched continuous control's .1242 (Figure 4). Several menus therefore improve the response relationships beyond matched training-state support.

The benefit extends to the held-out Reynolds numbers. Across Re100/170, effect nRMSE is .1595 for matched continuous coverage, .1017 for full-background coverage and .09965 for random Boolean coverage. Figure 4b separates interpolation from extrapolation and shows the higher remaining error at Re60. These descriptive comparisons condition on four fixed bases, with fits and histories nested within perturbation draw.

Halving the future time step on four prespecified sources preserves the coverage benefit while changing the ordering among Boolean menus. The matched controls reuse an inspected evaluation panel, so they establish coverage gains in this setting rather than prospective selection of a best menu. Numerical sensitivity and diagnostic-information costs are detailed in Appendix D.

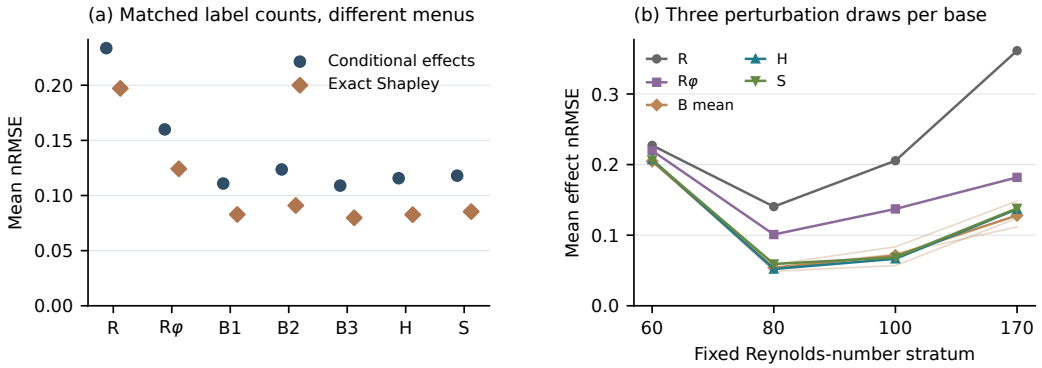


Figure 4: **Action coverage adds to the benefit of matched training support.** (a) Effect and exact-Shapley errors on twelve perturbation draws from four fixed bases. Rφ matches training parents/histories while retaining R's coefficients. B1–B3 are fixed random Boolean menus. (b) Re60/80 are trained values, Re100 tests interpolation and Re170 extrapolation. Pale lines retain individual B menus, whose errors are averaged for B mean. Histories and fits are nested within draw. Each arm uses 196 labels and the same update budget.

4.3 Why the executable question and the response check both matter

The learning studies depend on first asking the same question of model and reference. In transport and Burgers dynamics, we compare executed forcing with spatial replacement of the full-control field over 256 combinations of eight programs. Replacement fills unselected coordinates with zeros, training means or a Gaussian conditional completion supplied with the exact full-field mass. For transport with a Fourier neural operator (FNO), parent-normalized attribution error is .0719 for executable queries, .187 for conditional completion and 10.4 for zero replacement. Executable queries are closer in all four system–model combinations, including in raw units (Appendix C). Plausible field completion does not by itself recover the consequence of the prescribed forcing history.

Even with that question matched, exact attribution can reveal learned-response error. In a separate four-action 1D study, convolutional, Fourier neural operator and temporal convolutional models trained only on no-action examples achieve no-action future-field normalized mean-square error (NMSE) of $(6.67\text{–}7.78)\times10^{-7}$. Their local-energy Shapley errors are nevertheless 3.72%, 8.20% and 3.73%. Action-vertex and interior-amplitude training lowers these to .070%, .136% and .100% on the same 32-parent cohort, with the attribution rule fixed (Appendix Figure 5a). This cross-architecture comparison changes training coverage and costs, unlike the matched controls above.

Smoke transport shows the fidelity gap in two dimensions. On 40 trajectories with eight force programs, exact-Shapley error ranges from 41.80–53.21% for FNO and 79.10–92.05% for U-Net across five fitted checkpoints per architecture. The zero-allocation baseline has 57.96% error on this scale. FNO improves on that baseline and U-Net does worse despite exact enumeration of all 256 configurations (Appendix Figure 5b).

4.4 Field improvement does not ensure response improvement in smoke

Feedback in smoke transport does not recover the gains seen in conduction and cylinder flow. On 50 new trajectories, all fixed-400-update continuations have higher mean effect nRMSE than the starting model (.895–.951 versus .893), despite falling training losses and improved fields on inspected training trajectories. A separate ROI-objective test on 24 new trajectories gives favorable response-supervision point estimates, but neither primary interval excludes zero and other prediction metrics worsen. These results show why field gains cannot substitute for held-out relationship checks (Appendix G).

4.5 Executable predictions support physical hypothesis tests

The shared interface also makes a proposed physical relationship testable. In a separate 1D study, two fixed model-based rankings nominate cases where an action changes from increasing to decreasing local energy as control amplitude rises. All 24 distinct nominations on 22 parents retain the sign change under the performed time-step checks. In one preselected case, the magnitude of the negative overlap between the added and background fields exceeds the added field's positive energy contribution. Total local energy remains positive (Appendix Figure 7). Initial and repaired models nominate the same event sets, so the test does not establish improved discovery yield. Appendix F includes cylinder hypotheses, rejected interpretations and missed events.

5 Related Work

Specifying the experimental target. Removal-based explanations distinguish what is removed, how the model is queried and how effects are summarized (Covert et al., 2021). Conditional and interventional feature games answer different questions (Chen et al., 2020), while on-manifold approaches use conditional distributions to respect observed dependencies (Frye et al., 2021). Causal Shapley values introduce intervention assumptions (Heskes et al., 2020), and physical perturbation studies evolve altered states (Encinar & Jiménez, 2023). Executable games specify forcing, boundary or material programs so that each queried configuration has an explicit experiment shared by model and reference.

Comparing learned and reference relationships. Interventionally consistent surrogates learn simulator agreement under matched interventions through causal abstraction and learned intervention maps (Dyer et al., 2024). Data-faithful attribution addresses the data-generating process under unobserved confounding (Sun et al., 2024). We build on intervention agreement by locating errors in specific action backgrounds and testing feedback that improves them. Shapley–Taylor and Faith-Shap allocate interactions within a given value function (Sundararajan et al., 2020; Tsai et al., 2023). Such summaries are compatible with executable games. Here we hold the allocation rule fixed to study disagreement in the underlying learned responses.

Choosing feedback and supervision. Explanation-guided learning incorporates scientist feedback and gradient constraints (Ross et al., 2017; Schramowski et al., 2020). Sobolev training supervises derivatives alongside values (Czarnecki et al., 2017), and sensitivity-constrained operators can use finite-difference references (Behroozi et al., 2025). Active learning for neural PDE solvers selects simulation inputs to reduce solver calls (Musekamp et al., 2025). Our controls distinguish action coverage from how existing labels supervise the model. Conduction compares finite-amplitude differences with absolute temperatures on identical endpoint fields, then tests background dependence beyond the directly supervised singleton contrasts. Appendix A expands these distinctions.

6 Discussion and Conclusion

An executable explanation connects what a model has learned to a change that can be checked and used for feedback. Our studies show why this connection matters: accurate fields and averaged attributions can conceal incorrect conditional effects, while checking those effects identifies a concrete learning target. Action coverage improves cylinder responses after matching training support, and response supervision improves conduction effects and attributions from the same endpoint labels.

Endpoint labels contain both outcomes and relations between outcomes, but the objective need not teach them equally well. Emphasizing singleton differences also reduces mean error in predicting background dependence in new materials, even though this relationship is never directly supervised. In smoke, improved field fitting alone does not establish response improvement.

These conclusions concern specified action families, readouts and numerical references. Researchers choose feedback, with numerical and sampling scope documented in Appendices B–G. Preserving the action question from explanation through feedback lets us identify a learned relationship, improve it and determine where that improvement extends.

AI USE STATEMENT

We used large language models to assist with writing and language polishing, retrieve and examine related literature, support research ideation and execution, and draft manuscript sections. Research assistance included discussing the conceptual framework and mathematical exposition, proposing and refining experimental designs, implementing and debugging code, running and analyzing numerical experiments, and preparing figures and tables. Writing assistance included revising structure, improving readability and reference formatting. Reported numerical results come from the specified simulations, training runs and saved outputs, not from language-model-generated measurements. The human authors take responsibility for the final text, references, methods, results and scientific interpretations, including all AI-assisted content.

REPRODUCIBILITY STATEMENT

Appendix B specifies the model interfaces, objectives and statistical units, and Appendices C–G detail the individual studies. The anonymous supplementary code archive includes a README, environment requirements, compact numerical results and a single-command saved-result replay. Its core conduction recipe provides configuration-driven material generation, baseline and P/A/T training, checkpoint evaluation and statistical recomputation. A separate small CPU example demonstrates the pipeline. Scalar games and saved predictions support the reported analyses without rerunning simulations or training. The README distinguishes these replays from the from-scratch conduction commands and documents their compute requirements. Large checkpoints and raw field datasets are omitted; the conduction generator recreates the required training fields. The other systems include analysis assets and documented configurations, but not complete end-to-end retraining packages.

REFERENCES

Jannis Becktepe, Aleksandra Franz, Nils Thuerey, and Sebastian Peitz. Plug-and-play benchmarking of reinforcement learning algorithms for large-scale flow control. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62198.

Abdolmehdi Behroozi, Chaopeng Shen, and Daniel Kifer. Sensitivity-constrained Fourier neural operators for forward and inverse problems in parametric differential equations. In *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 12916–12937, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/227b19598f79ed838b01933b9a6ace41-Abstract-Conference.html.

Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data?, 2020. URL https://arxiv.org/abs/2006.16234. arXiv preprint arXiv:2006.16234.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://jmlr.org/papers/v22/20-1316.html.

Andrés Cremades, Sergio Hoyas, Rahul Deshpande, Pedro Quintero, Martin Lellep, Will Junghoon Lee, Jason P. Monty, Nicholas Hutchins, Moritz Linkmann, Ivan Marusic, and Ricardo Vinuesa. Identifying regions of importance in wall-bounded turbulence through explainable deep learning. *Nature Communications*, 15:3864, 2024. doi: 10.1038/s41467-024-47954-6. URL https://doi.org/10.1038/s41467-024-47954-6.

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. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/758a06618c69880a6cee5314ee42d52f-Abstract.html.

Joel Dyer, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. Interventionally consistent surrogates for complex simulation models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21814–21841, 2024. doi: 10.52202/079017-0686.

URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/26b8e3dc3a21fcd660d80c63b767f324-Abstract-Conference.html.

Miguel P. Encinar and Javier Jiménez. Identifying causally significant features in three-dimensional isotropic turbulence. *Journal of Fluid Mechanics*, 965:A20, 2023. doi: 10.1017/jfm.2023.407. URL https://doi.org/10.1017/jfm.2023.407.

Aleksandra Franz, Hao Wei, Luca Guastoni, and Nils Thuerey. PICT–a differentiable, GPU-accelerated multi-block PISO solver for simulation-coupled learning tasks in fluid dynamics. *Journal of Computational Physics*, 544:114433, January 2026. ISSN 0021-9991. doi: 10.1016/j.jcp.2025.114433. URL https://doi.org/10.1016/j.jcp.2025.114433.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1229–1239. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d770c496aa3da6d2c3f2bd19e7b9d6b-Abstract.html.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OPyWRrcjVQw.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=dPSTDbGtBY.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4778–4789. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

Philipp Holl, Vladlen Koltun, and Nils Thuerey. Learning to control PDEs with differentiable physics. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HyeSin4FPB.

Zongyi Li, Nikola B. 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. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Daniel Musekamp, Marimuthu Kalimuthu, David Holzmüller, Makoto Takamoto, and Mathias Niepert. Active learning for neural PDE solvers. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 38777–38825, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/60960ad78868fce5c165295fbd895060-Paper-Conference.pdf.

Art B. Owen. Sobol' indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2 (1):245–251, 2014. doi: 10.1137/130936233. URL https://epubs.siam.org/doi/10.1137/130936233.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5, Part 2):64–79, 1972. doi: 10.1287/mnsc.18.5.64.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, Cham, 2015. Springer. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://www.ijcai.org/proceedings/2017/371.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2:476–486, 2020. doi: 10.1038/s42256-020-0212-3. URL https://www.nature.com/articles/s42256-020-0212-3.

Eunhye Song, Barry L. Nelson, and Jeremy Staum. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083, 2016. doi: 10.1137/15M1048070. URL https://doi.org/10.1137/15M1048070.

Qiheng Sun, Haocheng Xia, and Jinfei Liu. Data-faithful feature attribution: Mitigating unobservable confounders via instrumental variables. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 44935–44964. Curran Associates, Inc., 2024. doi: 10.52202/079017-1428. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4f84e81c0a4eed2024cebcfb8f9d6e7f-Paper-Conference.pdf.

Mukund Sundararajan and Amir Najmi. The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278. PMLR, 2020. URL https://proceedings.mlr.press/v119/sundararajan20b.html.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268. PMLR, 2020. URL https://proceedings.mlr.press/v119/sundararajan20a.html.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-Shap: The faithful Shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023. URL https://jmlr.org/papers/v24/22-0202.html.

Geoff Ward, Sean Kamkar, and Jay Budzik. An exploration of the influence of path choice in game-theoretic attribution algorithms, 2020. URL https://arxiv.org/abs/2007.04169. arXiv preprint; version 2, 26 February 2021.

Supplementary guide. The supplement follows the paper's scientific questions rather than the order in which experiments were developed. Appendix A positions the closest methods; Appendix B defines measurements and mathematical details. Appendix C gives target and model diagnostics, Appendices D and E give the two learning studies, Appendix F reports physical hypothesis tests, and Appendix G examines repair transfer. The comparisons and operational definitions needed to interpret the main claims are included in these chapters.

A Extended Related Work: Which Question Is Shared?

Table 5 distinguishes approaches by the object changed and the question answered. An evolved initial-state perturbation and a source control can both define meaningful experiments while answering different questions.

Table 5: Questions and assumptions in related approaches. "Reference" denotes an independently executed numerical experiment for the stated action. Rows describe the cited formulation, not every possible adaptation.

Line of work	Changed object / assumed structure	Quantity answered	Relevant boundary
Removal explanation (Covert et al., 2021)	Input features under a specified removal operation	Change in model behaviour, summarized by an explanation rule	Physical meaning depends on the removal question
Flow-region attribution and perturbation (Cremades et al., 2024; Encinar & Jiménez, 2023)	Observed flow regions, or a localized initial condition then evolved	Predictive importance or dynamical significance of the perturbed region	Does not by itself define a forcing or wall-control action
Causal/asymmetric Shapley (Heskes et al., 2020; Frye et al., 2020)	Features under causal intervention assumptions or permissible orders	Contribution to a model prediction	We execute supplied programs rather than infer causal structure
Variance sensitivity (Owen, 2014; Song et al., 2016)	Uncertain inputs under a distribution	Share of output variance	Global variance, not a signed effect at one parent
Intervention-consistent surrogate (Dyer et al., 2024)	Simulator interventions represented in a surrogate	Simulator–surrogate agreement under interventions	Shared motivation; we also compare attribution and conditional effects
Data-faithful attribution (Sun et al., 2024)	Features under instrumental-variable assumptions	Attribution consistent with data generation	We use executable references, not hidden-confounder identification
Guided learning (Behroozi et al., 2025; Ross et al., 2017; Schramowski et al., 2020)	Parameter sensitivity or annotated explanation constraints	Better derivative behaviour or corrected model reliance	We test finite action contrasts with matched endpoint-information and optimization controls
This study	Executable source or material programs on a shared parent	Response, conditional effect and attribution versus reference	A common intervention and measurement are required

Experimental and attribution targets. Removal and causal Shapley approaches specify the value function through a removal operation or causal assumptions (Covert et al., 2021; Heskes et al., 2020). We supply executable source, boundary or material programs. Our completion comparison asks whether a reconstructed field substitutes for a prescribed source action; field plausibility alone does not specify that action's program or history. It does not dispute the validity of evolved state-perturbation experiments (Encinar & Jiménez, 2023).

Resolution of the matched comparison. Interventional consistency checks simulator–surrogate agreement (Dyer et al., 2024), and model/data fidelity is an established distinction (Sun et al., 2024). We examine both attribution across backgrounds and effects at a specified background. Exact enumeration removes allocation-estimation error, but averaging can still conceal conditional errors. Both summaries organize the same endpoint information; neither adds new information to it.

Feedback targets and matched controls. Explanation feedback includes augmentation and gradient constraints (Schramowski et al., 2020); sensitivity-constrained operators can estimate derivatives by finite differences (Behroozi et al., 2025). We target finite observable changes at prescribed physical amplitudes. Cylinder experiments test parent/history and action-distribution coverage under one ordinary objective. Conduction experiments compare response and absolute-observable supervision from the same labels, with optimization controls. These comparisons distinguish label choice, supervision and training budget; separate reference experiments test proposed physical consequences.

B Experimental Units, Measurement and Mathematical Details

B.1 Experimental units and execution interfaces

A parent specifies the physical starting condition and relevant history; all branches share its measurement window and readout. Table 6 distinguishes physical units from repeated coalitions and fitted models. Study-specific settings appear in Appendices C–G.

Table 6: Evaluation panels and dependence. Panel sizes cannot be summed into an overall sample size.

Panel	Physical units and repeated measurements
1D completion	40 parents per system, shared by architectures and five checkpoint seeds
1D coverage	Separate 32- and 64-parent cohorts; three architectures and five seeds share training data within regime
Smoke diagnosis	40 trajectory parents; ten checkpoints and 256 subsets per parent
Cylinder coverage	Four saved velocity/pressure bases, one per Re; three native perturbation draws per base, with two histories and three fits nested within draw
Conduction	One inspected 16-material panel and four disjoint 64-material panels (fixed-update, long-budget, modulation, finite-weight confirmation), each paired within material across three fixed fits; separate 16-material development for weight selection
Smoke repair	Separate 50- and 24-trajectory panels, with fixed fits paired within trajectory
1D hypotheses	24 events on 22 parents; events share parents and controlled amplitude grids
Cylinder hypotheses	Four new mother states; future-step settings share each starting history

For post-control field prediction, execution and measurement take the form

$$x_p(S) = \Phi^{\mathrm{ctrl}}\left(p, \sum_{i\in S} \alpha_i b_i\right), \tag{7}$$

$$G_{p,\mathrm{ref}}(S) = Y\big(\Phi^{\mathrm{fore}}(x_p(S))\big), \qquad G_{p,\theta}(S) = Y(f_\theta(H_p(S))), \tag{8}$$

where $H_p(S)$ contains the available controlled state and history. Smoke uses this executed-prefix interface. Cylinder models directly predict null-relative force responses from pre-control information and coefficients; conduction models predict equilibrium temperature for the changed conductivity. The shared requirement is the intervention and measurement, not the input representation. Local energy and downstream smoke are regional readouts, not conservation tests; lift RMS is unsigned. Each study defines these quantities and its reference checks where first used.

B.2 Errors, normalization and uncertainty

Write $e_p(S) = v_{p,\theta}(S) - v_{p,\mathrm{ref}}(S)$. Response, effect, pair and attribution RMSE respectively concern e_p, $d_{i,\theta} - d_{i,\mathrm{ref}}$, $I_{ij,\theta} - I_{ij,\mathrm{ref}}$ and $\phi_\theta - \phi_{\mathrm{ref}}$. Absolute-endpoint and future-field errors are separate

targets. The 1D coverage study averages action-coordinate RMS within parents and divides by the common cohort scale

$$s_q = \left[\frac{1}{P(2^m - 1)} \sum_p \sum_{S \neq \emptyset} v_{p,\mathrm{ref},q}(S)^2 \right]^{1/2}. \tag{9}$$

Smoke diagnosis instead divides each parent's action RMS by its RMS over 256 null-centered reference vertices before averaging parents. Completion uses parent-scaled ratios, with companion raw local-energy errors. Cylinder uses four fixed training-derived output scales and equal-Re aggregation; conduction uses temperature-unit effect/Shapley RMSE and explicitly identified relative or raw field error. Smoke confirmations retain both normalized response and raw ROI-effect metrics. These scales are not interchangeable.

Bold and underlined error means rank first and second within each stated comparison; paired intervals report uncertainty in the specified contrasts.

The 75-model coverage analysis jointly resamples parent and seed IDs across methods, fixes architecture strata and recomputes the denominator, conditional on saved training data. Cylinder confirmation averages histories/fits within source, then resamples sources within Re. Conduction new64 and smoke repair resample whole materials and trajectories, respectively, preserving method pairing and fixed fits. Study chapters distinguish adjusted primary from descriptive secondary intervals. Edges are not independent physical replicates: a six-action cylinder game has 64 endpoints and 192 edges per parent/history.

B.3 Attribution identities and response errors

For $e = v_\theta - v_{\mathrm{ref}}$, Shapley linearity gives $\phi(v_\theta) - \phi(v_{\mathrm{ref}}) = \phi(e)$. Its nonnegative marginal weights sum to one, and every marginal error is bounded by $2\|e\|_\infty$, giving

$$\|\phi(v_\theta) - \phi(v_{\mathrm{ref}})\|_\infty \le 2\|e\|_\infty, \qquad \|\phi(v_\theta) - \phi(v_{\mathrm{ref}})\|_2 \le 2\sqrt{m}\|e\|_\infty. \tag{10}$$

Raw endpoint errors bounded by δ also give the coordinate bound 2δ, since the null error cancels. The reverse fails: in endpoint order $(\emptyset, \{1\}, \{2\}, \{1,2\})$, error $(0, a, a, 0)$ has zero Shapley allocation but nonzero conditional errors. Moreover, replacing $(0,1,1,0)$ by $(0, .4, .4, .8)$ lowers every edge-error magnitude from 1 to .4 while raising both Shapley errors from 0 to .4. Improvements in these two summaries are therefore distinct findings, but not independent replications.

For a continuously differentiable amplitude response $y : [0,1]^m \to \mathbb{R}$, $y(0) = 0$, define $v(S) = y(\mathbf{1}_S)$, its multilinear extension $f_v(z) = \sum_{S \subseteq N} v(S) \prod_{i \in S} z_i \prod_{i \notin S}(1 - z_i)$, and diagonal IG $\psi_i(y) = \int_0^1 \partial_i y(t\mathbf{1})\,\mathrm{d}t$ (Sundararajan et al., 2017). The multilinear-extension identity is due to Owen (1972). To compare this vertex allocation with path attribution (Sundararajan & Najmi, 2020), we derive below

$$\phi_i(v) = \int_0^1 \partial_i f_v(t\mathbf{1})\,\mathrm{d}t, \qquad \psi(y) - \phi(v) = \int_0^1 \nabla(y - f_v)(t\mathbf{1})\,\mathrm{d}t, \tag{11}$$

$$\psi_i\left(c \prod_{j \in T} z_j^{a_j}\right) - \phi_i(v) = c\left(\frac{a_i}{k} - \frac{\mathbf{1}[i \in T]}{|T|}\right), \qquad k = \sum_{j \in T} a_j. \tag{12}$$

In the second line, v is the displayed monomial's vertex game, $a_j \ge 1$ on T and $a_i = 0$ outside it. To verify the identities, use dividends $a_T = \sum_{R \subseteq T} (-1)^{|T|-|R|} v(R)$ and expand $f_v = \sum_{T \neq \emptyset} a_T \prod_{i \in T} z_i$: integrating each diagonal derivative assigns $a_T/|T|$. A monomial's vertex game shares c equally over T, whereas its IG shares ca_i/k. Thus degree-at-most-two polynomials have zero rule gap, consistent with Ward et al. (2020). All multilinear interactions also have zero gap by Eq. (11); $x^2 y$ has gap $(1/6, -1/6)$, canceled by xy^2. Interaction alone neither determines the gap nor measures physical error. Same-rule IG disagreement is instead $\int_0^1 \partial_i(y_\theta - y_{\mathrm{ref}})(t\mathbf{1})\,\mathrm{d}t$: vertex agreement controls Shapley but need not control diagonal gradients.

B.4 Contrast supervision from endpoint labels

An edge matrix B has coefficients $(-1, +1)$ at $(S, S+i)$; a square matrix C has $(+1, -1, -1, +1)$ at $(S, S+i, S+j, S+i+j)$. Losses $\|Be\|^2$ and $\|Ce\|^2$, divided by the contrast counts, reuse

endpoint residuals e. Full-LOO restricts edges to the all-actions vertex and is not an interaction loss. For normalized Boolean Walsh characters, $e(z) = \sum_T \widehat{e}_T \chi_T(z)$, uniform averages satisfy

$$\mathbb{E}_{i,S}|D_i e(S)|^2 = \frac{4}{m}\sum_T |T|\widehat{e}_T^2, \qquad \mathbb{E}_{i<j,S}|D_{ij}e(S)|^2 = \frac{16}{\binom{m}{2}}\sum_T \binom{|T|}{2}\widehat{e}_T^2. \tag{13}$$

A difference annihilates characters missing its coordinate and otherwise multiplies them by two; orthogonality removes cross terms. These identities describe loss weighting; the controls below test its learning benefit.

B.5 Optimization and output-supervision controls

A complete-label 1D study used 128 training parents (all 2,048 endpoints), 32 validation parents and three paired ResCNN seeds. Six objectives took 4,800 updates; longer ordinary fits took 5,148/5,177/5,146 updates to match the longest observed per-seed time. All 21 checkpoints preceded testing on 64 new parents at five amplitudes. At amplitude one, ordinary/all-edge/pair/ longer-ordinary edge errors were $.182148/.161122/.121315/.168344\%$ of $s_E = .2596097553$. All five all-edge-minus-longer-ordinary intervals cross zero; field and gradient-energy errors worsen in all fifteen paired method–seed comparisons with that control.

A second study compares pair and direct-output objectives on labels acquired by selectors A (ensemble endpoint uncertainty) and D (predicted squared contrast residual plus contrast variance per missing endpoint). Three data draws each use 512 initial plus 512 acquired endpoints, a separate 512-endpoint/ 32-parent development panel, and common initial checkpoints across final arms. L_R combines normalized field MSE and .1 times local-energy endpoint MSE. For $Q \le 64$ completed squares per update, write $e_{qk,o} = (\widehat{G}_{qk,o} - G_{qk,o})/s_o$, $o \in \{V, E\}$, with corners $k = (S, S+i, S+j, S+i+j)$, and $V = \sum_x w_x u_x$, $E = \frac{1}{2}\sum_x w_x u_x^2$ using the readout weights in Appendix C.2. The objectives are

$$L_{\text{pair}} = L_R + \frac{1}{2Q}\sum_{q,o}(e_{q0,o} - e_{q1,o} - e_{q2,o} + e_{q3,o})^2, \tag{14}$$

$$L_{\text{Direct}} = L_R + \frac{4}{8Q}\sum_{q,k,o} e_{qk,o}^2. \tag{15}$$

Two-candidate historical development searches selected the weights; they matched candidate counts, not search histories or gradient magnitudes. Pair/Direct take 4,800 updates and R-long 8,593, assigned from measured ordinary/pair update times of 5.4524/9.7612 ms; Direct is .388% faster than pair. Development energy MSE, then field MSE, selects among 26 checkpoints per arm. The test has 64 new parents and 32 edges each. Energy edge RMSE pools squared errors over edges, parents and three draws; 10,000 paired crossed draw/parent resamples give descriptive, unadjusted 95% intervals. Full selector, normalization, optimizer and selection settings are in the submitted code package's `README_v80_1d_controls.md`.

Table 7: Same-label 1D controls. Positive edge reductions favor pair; positive field changes are adverse. The prespecified rule required $\ge 2\%$ edge improvement and $\le 5\%$ field increase. Brackets are descriptive 95% crossed-bootstrap intervals.

Pair comparison	Edge reduction (%)	Field change (%)
D vs. R-long	$4.70\ [-2.73, +9.14]$	$+9.53$
A vs. R-long	$2.13\ [-6.24, +6.26]$	$+9.39$
D vs. Direct	$17.57\ [+4.66, +25.41]$	-5.12
A vs. Direct	$12.69\ [+3.39, +16.67]$	-6.23

In Table 7, both pair–R-long intervals cross zero and pair exceeds the field-cost criterion. Pair improves edge and field errors over Direct in all three draws. The gain therefore depends on the objective and optimization comparator.

B.6 Information use and measured costs

Table 8 reports measured costs. Information budgets count distinct revealed endpoints. A four-corner contrast needs zero to four new labels; reuse removes new reference queries, not training or diagnostic cost. Cylinder matched arms use 196 endpoints; the 360-endpoint control includes labels consulted during diagnosis. Four-, six- and eight-action games enumerate 16, 64 and 256 endpoints. Selected contrasts can use fewer, but large-game simulation savings are not established. Numerical checks cover the stated resolutions and histories, not continuum-error certificates.

Table 8: Measured costs. Timings have different scopes and are not end-to-end savings estimates; conduction timings use two DGX CPU threads. Long-budget arm times remain beside their comparison in Appendix E.2.

Study	Revealed fields / measured time
1D complete-label	2,048 train, 512 nominal validation, 1,024 additional-amplitude development, 5,120 confirmation and 160 refinement fields, plus warmup/control preparation; 54 development/final fits: 2,277.108 s (final 21: 1,417.83 s)
Cylinder	Twelve-source generation: 50,347.95 s wall, including 18,396.34 s allocation wait (31,951.61 s active); selected Re170 eight-branch refinement: 292.79 s solver / 296.04 s driver. Nested branch timers are not additive
Conduction original	656 solves (648 at 32^2, eight at 64^2), excluding eight unit tests; 256 form the inspected 16-material games. Four original fits: 598.50 s; two further B fits: 404.031 s; four P/T continuations: 399.526 s; three A continuations: 298.196 s
Conduction fresh64	1,024 solves and twelve-model evaluation: 32.77 s wall, excluding transfer and summaries

C What the Matched Diagnosis Measures

Model and reference answer the same action game. Exact Shapley enumeration removes allocation-estimation error, allowing discrepancies in learned responses to be examined at the units and scales in Appendix B.

C.1 When state completion does not answer the prescribed action

The eight-action transport/Burgers comparison uses 40 shared parents, two architectures and five checkpoints per system. Every allocation targets the parent's executable-game reference. Static replacement starts from the full-control end field and retains selected projected supports and coordinates outside their union, including overlaps. Each support is the smallest grid set containing 95% of an actuator's time-integrated squared amplitude. Gaussian completion conditions on retained coordinates and exact full-coalition spatial mass. Its mean/covariance use 512 training fields per system, with ridge 10^{-4} times mean marginal variance. This favorable mass constraint does not supply the source program's temporal history.

Allocation error $e_{p,c}$ is RMS over eight actions; s_p is RMS over 256 null-centered reference vertices. Table 9 reports means of the ratio $e_{p,c}/s_p$ (not percentages; the 10^{-12} floor is inactive) and paired raw local-energy differences. Its 95% intervals use 5,000 crossed resamples of five checkpoint seeds and 40 parents.

Completion improves on zero and train-mean replacement, but executable queries have lower error in all four cells on both scales. Ranking need not improve with magnitude: Burgers/FNO Gaussian-minus-zero absolute-Spearman agreement is $-.3463$ ($[-.5405, -.1559]$). This comparison combines query mismatch, completion approximation and learned response error; it neither isolates those contributions nor evaluates other state-perturbation questions.

C.2 Exact attribution can expose an inaccurate learned response

The reference solves $\partial_t u = -.35\partial_x u - \beta u \partial_x u + .04\partial_{xx} u + f$ on periodic $[0, 2\pi)$: 128 points, Fourier derivatives, two-thirds nonlinear cutoff and RK4 ($\Delta t = .002$). Zero-mean parents combine eight

Table 9: Completion versus executable queries. Query entries are mean parent-scaled allocation-error ratios; zero/train-mean replacements share one column. Gaussian-minus-executable gaps give their mean and 95% interval on the next line. Raw gaps use local-energy units. The same 40 parents and five checkpoint seeds underlie all columns; Gaussian completion receives exact full-coalition mass.

System / model	Executable	Gaussian	Zero / mean	Scaled gap	Raw gap
Transport / FNO	**.07194**	.18687	10.4 / 10.4	.11493	.002407
				[.10372, .12498]	[.001926, .002872]
Transport / TCN	**.07786**	.19707	11.6 / 11.6	.11920	.002423
				[.10759, .13050]	[.001971, .002882]
Burgers / FNO	**.12158**	1.10347	14.6 / 14.5	.98189	.008738
				[.63359, 1.38706]	[.006030, .011764]
Burgers / TCN	**.11183**	1.38905	14.9 / 14.9	1.27722	.009977
				[.75561, 1.97089]	[.007135, .013257]

random Fourier modes (RMS .6) and .2 time units of nonlinear warmup. Four stationary compensated Gaussian sources are centered at $\pi + (-1.5, -.5, .5, 1.5)$, with inner/outer widths $.30/.60$, zero mean and spatial RMS .8 before amplitude scaling. Control lasts .8 at $\beta = 0$; source-free forecasting lasts .2 at $\beta = .75$. Predictors receive the controlled field and preserve its mean through a compensated residual. All branches share local-energy readout $E(u) = \frac{1}{2}\sum_x w_x u_x^2$, with normalized Gaussian weights centered at $\pi + .25$, width .5, and the same forecast window.

Five seeds and five coverage regimes yield 75 predictors: width-32, three-block ResCNN (18,849 parameters); width-24, three-layer FNO with 16 modes (57,169) (Li et al., 2021); and width-24, five-block dilated TCN (17,689). AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, batch 256, a plateau scheduler, at most 300 epochs and validation stopping with patience 60 after epoch 50. All regimes have 2,048 training and 512 validation rows, but different training-parent counts: 2,048 factual-only, 128 vertices-only and 64 for the other regimes. Mixed coverage uses 16 vertices, eight interior and eight diagonal amplitudes per parent; continuous-only uses 32 interior amplitudes. Sensitivity adds four tangent directions to mixed at normalized weight one and about five times its training cost. Seeds share training data within regime; test cohorts A (32 parents) and B (64) are disjoint from training.

Table 10 retains all regimes and separates same-rule model–reference errors from the model's IG–Shapley gap. Accurate factual fields can coexist with inaccurate action attribution. Factual-to-mixed coverage lowers cohort-A Shapley error by an architecture-balanced 5.114 percentage points (parent/seed bootstrap interval $[4.168, 6.193]$); cohort B shows the same pattern. This diagnoses coverage at the stated parent counts and costs, not equal-budget learning. Continuous-only coverage improves IG while leaving larger Shapley errors than vertex-containing regimes. The reference energy IG–Shapley gap is about 10.56% of cohort A's scale, larger than the model–reference errors after intervention coverage.

A matched-coverage follow-up gives mixed evidence for diagonal versus interior amplitudes: velocity IG error changes by $-.03431$ percentage points for ResCNN ($[-.06151, -.00732]$) and $+.20599$ for FNO ($[-.01499, .44273]$). Derivative supervision improves nominal energy IG error at fixed updates but loses to time-matched ordinary training.

C.3 The same rule in two-dimensional smoke flow

The Navier–Stokes reference has buoyancy .5, viscosity .01, domain $[0, 32]^2$ and a 128^2 grid. PDEArena data (Gupta & Brandstetter, 2023) are branched with PhiFlow (Holl et al., 2020) (version 2.1.0), step 1.5 and every-fourth-step saves. Eight interior streamfunction modes define pressure-projected curl forces at coefficient .01, checked for divergence, walls, finite values, rank and quadratic budget .0064. Control takes factual F_6 to C_7; input $[F_4, F_5, F_6, C_7]$ forecasts smoke and two velocities through C_8, C_9, C_{10}. The readout integrates C_{10} smoke over fractional bounds $(.70, .92, .25, .75)$, with cell area .0625. Each evaluator subtracts its own no-action outcome.

All 256 coalitions are evaluated on 40 trajectories for five factual checkpoints each of local FNO (Li et al., 2021) and U-Net (Ronneberger et al., 2015) implementations, selected without intervention

Table 10: Coverage cohorts A (32 parents) and B (64). Shapley and IG errors compare the same rule in model and reference; entries are mean ± sample SD across five fitted seeds. Gap is the model's mean IG–Shapley gap. All action quantities are percentages of the cohort response scale. Factual-only no-action field NMSE appears below on its separate squared-field scale.

		Cohort A			Cohort B		
Architecture	Coverage	Shapley	IG	Gap	Shapley	IG	Gap
CNN	Factual	3.722 ± 0.183	0.748 ± 0.047	7.60	3.278 ± 0.156	0.692 ± 0.048	7.45
	Vertices	0.065 ± 0.016	0.095 ± 0.016	10.57	0.067 ± 0.016	0.091 ± 0.011	10.07
	Mixed	0.070 ± 0.004	0.074 ± 0.010	10.55	0.079 ± 0.010	0.074 ± 0.006	10.07
	Interior	0.182 ± 0.043	0.071 ± 0.019	10.45	0.183 ± 0.042	0.069 ± 0.014	9.98
	Sensitivity	0.056 ± 0.012	0.056 ± 0.009	10.55	0.055 ± 0.010	0.051 ± 0.007	10.07
FNO	Factual	8.195 ± 0.687	2.655 ± 0.462	5.34	6.809 ± 0.550	2.057 ± 0.344	5.37
	Vertices	0.123 ± 0.015	0.211 ± 0.047	10.60	0.127 ± 0.008	0.198 ± 0.026	10.09
	Mixed	0.136 ± 0.022	0.150 ± 0.017	10.56	0.143 ± 0.013	0.151 ± 0.019	10.05
	Interior	0.309 ± 0.020	0.135 ± 0.010	10.36	0.297 ± 0.030	0.126 ± 0.008	9.88
	Sensitivity	0.094 ± 0.009	0.087 ± 0.014	10.55	0.099 ± 0.009	0.085 ± 0.007	10.06
TCN	Factual	3.729 ± 0.163	0.800 ± 0.106	7.66	3.283 ± 0.144	0.735 ± 0.088	7.47
	Vertices	0.089 ± 0.010	0.141 ± 0.016	10.57	0.087 ± 0.010	0.135 ± 0.026	10.05
	Mixed	0.100 ± 0.007	0.107 ± 0.019	10.55	0.100 ± 0.012	0.111 ± 0.016	10.05
	Interior	0.168 ± 0.032	0.100 ± 0.010	10.49	0.179 ± 0.022	0.108 ± 0.013	9.98
	Sensitivity	0.064 ± 0.013	0.062 ± 0.008	10.56	0.064 ± 0.014	0.060 ± 0.008	10.06

Factual-only no-action NMSE on cohort A: CNN $(7.78 \pm 1.18) \times 10^{-7}$, FNO $(6.67 \pm .985) \times 10^{-7}$, TCN $(6.85 \pm 1.41) \times 10^{-7}$ (mean ± seed SD). For each seed, mean squared field error is divided by mean squared reference field at the null endpoint, pooling all 32 parents and coordinates before taking the ratio. This differs from attribution error and relative L_2.

labels. All reported models run in float64, replacing an earlier inconsistent GPU float32 FNO execution. Ordinary field errors are about 10% on a separate field scale; Table 11 uses parent response scales.

Table 11: Smoke diagnosis: ranges across five checkpoints per architecture of mean parent-scaled errors over 40 shared trajectories, in percent. Ranges are not confidence intervals. Shapley/IG errors each compare the same rule in model and reference.

Metric (% parent response scale)	FNO	U-Net
Response RMSE	82.66–105.73	144.08–175.62
Same-rule Shapley error	41.80–53.21	79.10–92.05
Same-rule IG error	41.69–53.22	79.23–92.36
Estimated model IG–Shapley gap	.295–.334	.725–.915

The zero-response predictor has 100% response error; zero allocations have 57.96% mean Shapley error on the same scale. Same-rule discrepancies exceed the estimated reference IG–Shapley gap (mean $.9037\%$). The smaller model rule gaps have $K = 8 \to 16$ and finite-difference-step sensitivities reaching $.6334\%$ and 1.0064% on two development parents, so they are not uniformly resolved. Exact Shapley discrepancies are independent of this IG approximation. Tighter reference calculations cover those two parents, not all forty.

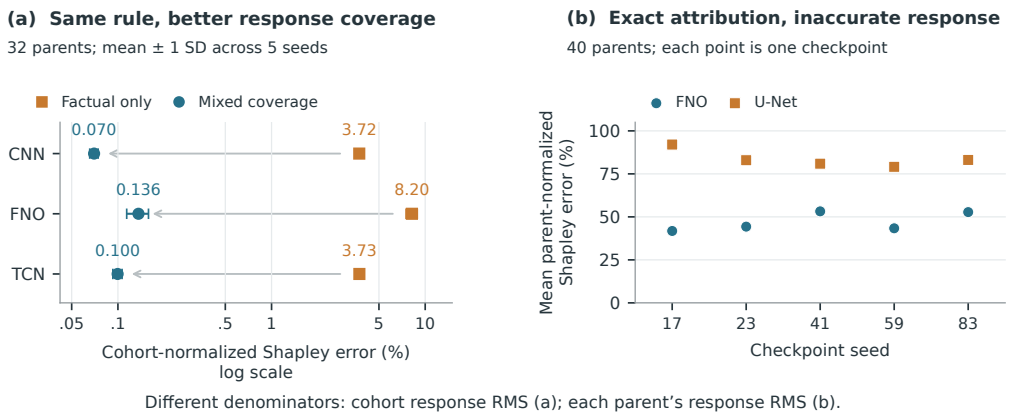


Figure 5: **Exact attribution changes with the learned response.** (a) Factual and mixed-coverage 1D means with one sample SD across five seeds, normalized by the cohort response scale (log axis). (b) Ten smoke checkpoints, each averaging parent-normalized errors over 40 trajectories. The panels use different normalization scales.

D Cylinder Action Coverage and Numerical Scope

This study asks how matching training parents and histories, then changing action-distribution coverage, affects response prediction on perturbations of four saved flow bases.

D.1 Actions, training menus and evaluation units

The reference uses FluidGym's rotating-cylinder environment (Becktepe et al., 2026) and its PISO-torch solver associated with PICT (Franz et al., 2026), at resolution setting 24 and nominal step .01. Six clamped cubic B-spline wall-motion modes receive a sine-squared envelope, separate discrete-L^2 normalization and rescaling by the peak of their full sum. A Boolean coalition sums selected modes with coefficient .08, without coalition-specific renormalization. Each branch applies 20 controlled commands of duration $.05D/U$, then 40 zero commands; the null uses the same parent with zero rotation. Overlapping modes make coalition energies unequal despite equal single-mode norms.

The CNN receives four pre-control 96×515 rendered-vorticity frames, six coefficients, Reynolds number and nominal phase. It predicts four null-relative scalars: post-control mean drag, signed mean lift, final-command signed lift and a rendered near-wake enstrophy proxy. The proxy is not a native-grid integral. Forces average PISO substeps, so final-command lift is a window average; post-control means use $1 < t \leq 3$. At each training Re60, 80, 120 and 160, 48 signed continuous candidates use antithetic Latin-hypercube coefficients, twelve per phase category and all six nonzero. An outcome-independent hash of group identity, seed and row reserves six per Re. The other 42 plus one full-positive endpoint form the common $4 \times 43 = 172$ labels.

Table 12 defines the 24-label additions. Rφ executes R's coefficient vectors from H's actual saved parent and history; it does not merely overwrite a phase input. S fixes masks $15, 30, 39, 51, 57, 60$ by cyclic adjacent-pair omission, with each action appearing four times. B1–B3 draw six distinct masks per Re uniformly without replacement from the 62 nonempty, nonfull masks; H/S masks remain eligible. Their menu seeds are 790251–790253, fixed before outcome inspection, with no best-menu selection. Each menu is crossed with fitted seeds 17, 23 and 41.

Continued arms start from paired seed-specific checkpoints with their original splits and normalization, and use fresh AdamW (learning rate .002, weight decay 10^{-4}), 1,200 updates and mean smooth-L1 loss ($\beta = .5$) over nonzero endpoints and four outputs. H-plus-loss adds a same-endpoint full-LOO term of weight .1. Every arm selects among step zero and every update on the same 48 ordinary Re140 validation responses, never evaluation-game errors. The 360-label information control includes $172 + 189 - 1$ consulted endpoints; one Re60 full endpoint is shared. S reuses six labels and adds 18

Table 12: Menus added to the common 172 endpoints at training Re60/80/120/160. H is full-minus-one coverage; S is balanced size-four coverage. Cardinality counts nonzero coefficients. Rφ/H/S/B share H's phase-1 parent and history. Waveform mean squares describe the wall input, not flow energy.

Menu	Coefficients	Active actions	Phase coverage	Mean square $\times 10^4$
R	Signed continuous	6	Four categories	$2.031, 2.660, 1.820$
Rφ	Same vectors as R	6	Category 1	$2.031, 2.660, 1.820$
H	Positive .08	5	Category 1	11.292
S	Positive .08	4	Category 1	8.482
B1	Positive .08	1–5 (mean 3.083)	Category 1	5.789
B2	Positive .08	1–5 (mean 2.875)	Category 1	4.995
B3	Positive .08	1–5 (mean 3.083)	Category 1	5.831

R/Rφ triples follow fit seeds 17, 23, 41. Common-set phase counts $(0, 1, 2, 3)$ are respectively $(45, 43, 40, 44)$, $(43, 44, 41, 44)$, $(42, 48, 40, 42)$; R additions are $(3, 9, 8, 4)$, $(5, 8, 7, 4)$, $(6, 4, 8, 6)$. All other additions have counts $(0, 24, 0, 0)$.

solves; Rφ/B require 130 distinct queries (52 reused, 78 new). These controls match labels within the 196-label arms, not total diagnosis cost.

The evaluation panel has twelve prespecified native velocity/pressure perturbation draws, three conditional on each saved Re60, 80, 100 and 170 base. Re60/80 occur in training; Re100 is an interpolation test and Re170 an extrapolation test in Reynolds number. Geometry and resolution are fixed. Each draw integrates 66–131 steps at .01, runs 120 zero-action recovery commands and a 240-command scan, then supplies opposite 20-command sine-squared histories of parameter $\pm.04$ followed by ten zero commands. Four input frames are spaced by $.05D/U$; each reference branch restores its source and replays the history before future control.

The inherited local peak–trough midpoint selector retains `phase=1`, without guaranteeing macroscopic cycle alignment. The historical Re60 anchor is near a positive lift peak, whereas both development anchors are near negative troughs (about 50.5% relative input L^2 distance); two new Re60 draws differ by 6.33%, versus 0.225–2.535% at other Re values. Sources were neither filtered nor rephased after inspection. The controls S, Rφ and B reuse this already inspected evaluation panel; they add no independent evaluation sources.

The 24 histories supply $12 \times 2 \times 64 = 1,536$ nominal endpoints and 192 conditional edges per game. Residuals use historical training-output RMS scales $(.00379150, .02297857, .01446194, .00123027)$ from 168 endpoints, in output order above. We take RMS across standardized outputs and evaluated edges per history/fit, average histories and fits within draw, then average draws within Re and the four strata equally. Responses use 64 null-centered vertices; Shapley errors use six exact allocations. Scores are means of draw-level RMSEs, not pooled RMSEs. Pointwise 95% percentile intervals use 20,000 paired whole-draw resamples within Re, conditional on saved bases, menus and fits. With only three draws per stratum, coverage is approximate; menus and fitted seeds do not add independent physical samples.

D.2 Parent/history matching and action-distribution coverage

Table 13 retains every training role and Boolean menu. Matching parents and histories lowers Rφ's effect error by 31.5% relative to R. B mean then lowers it by 28.4% relative to Rφ: paired difference $-.045465$ $[-.046432, -.044088]$, favorable in all twelve draws. B mean averages per-draw error metrics after history/fit averaging, never predictions. Against H its difference is $-.001189$ $[-.004425, .000824]$, with only four favorable draws; against S it is $-.003508$ $[-.004724, -.002158]$, with nine favorable draws. The intervals condition on these three realized menus, not uncertainty over unseen menu designs.

R and Rφ share signed coefficients spanning approximately $\pm.079967$, all six nonzero. H/S use five/four positive .08 coefficients; B spans one to five active actions. Table 12 gives their waveform mean squares over the 60 commands, which measure prescribed wall input rather than fluid dissipation or work. B versus Rφ jointly changes sign, amplitude, zero coefficients, cardinality and energy. The

Table 13: Nominal twelve-draw results, with histories and fitted seeds averaged within draw and equal weighting across four Re strata. All errors use fixed training-output scales. Continued arms share 1,200 updates; the 360-label arm includes extra consulted information. B mean averages errors from all three menus. Bold/underline rank R, Rφ, H, S and B mean within each error column.

Training role	Labels	Response	Effect	Shapley
Initial	172	.469	.289	.246
Ordinary continuation	172	.415	.239	.198
Matched coverage arms (196 labels)				
R	196	.401539	.233640	.197135
Rφ	196	.279727	.159940	.124185
H	196	**.224672**	<u>.115664</u>	**.082613**
S	196	<u>.229721</u>	.117982	.085378
B mean	196	.234606	**.114475**	<u>.084512</u>
Individual random menus and additional feedback controls				
B1	196	.232385	.110850	.082816
B2	196	.246956	.123608	.090944
B3	196	.224479	.108967	.079777
H + full-LOO loss	196	.224	.115069	.0821
All consulted information	360	.194	.082391	.0660

result supports action-distribution coverage beyond parent/history matching, without isolating any one of these mechanisms or a uniquely effective full-LOO selector.

H reduces all-edge error by 50.5% relative to R: $-.117976$ $[-.121731, -.111983]$, favorable in 11/12 draws. On the 120 internal edges, excluding null, full, singleton and leave-one-out endpoints, H/R is $.116658/.226831$: $-.110173$ $[-.115185, -.102121]$. Those edges exclude H's supplied relations but can touch S/B training-mask coordinates; they are not universally unsupervised. Centered-response and Shapley H$-$R differences are $-.176867$ $[-.193111, -.151540]$ and $-.114522$ $[-.116543, -.111899]$. All outputs improve on average, but H-plus-loss adds little ($.115069$ effect error), and the 360-label control reaches $.082391$. These comparisons establish neither a specialized-loss gain nor a total-reference-cost advantage.

Table 14: Nominal all-edge effect nRMSE by fixed Re stratum. Each entry averages three perturbation draws and their nested histories/fits. Re60/80 are training Re values; Re100/170 test interpolation/extrapolation. B1–B3 are menu repetitions, not additional sources. Bold/underline rank R, Rφ, H, S and B mean within each row.

	Matched coverage arms					Individual random menus		
Re	R	Rφ	H	S	B mean	B1	B2	B3
60	.2272	.2198	.2071	<u>.2063</u>	**.2047**	.2043	.2040	.2059
80	.1405	.1009	**.0521**	.0592	<u>.0539</u>	.0541	.0586	.0489
100	.2055	.1372	**.0665**	<u>.0685</u>	.0713	.0735	.0835	.0569
170	.3614	.1819	<u>.1369</u>	.1380	**.1280**	.1116	.1483	.1241

The breakdown in Table 14 shows that B mean beats H at Re60/170 but not Re80/100. Its nominal Shapley error is slightly higher: difference $.001899$ $[.000809, .002608]$. Similar aggregate effect errors therefore do not imply the same ranking by stratum or metric.

Balanced size-four coverage. S reduces nominal effect error by 49.50% versus R: $-.115658$ $[-.116977, -.114167]$. Its small difference from H is $+.002318$ $[-.002185, .004957]$, with H lower on 11/12 draws. Table 15 retains the paired intervals and output/fit heterogeneity. Nominal Shapley error favors H, whereas its finer-step ordering slightly reverses. An interval crossing zero establishes neither equivalence nor menu superiority; comparisons of S with H also change action cardinality and waveform energy. Full masks and source/output/fit results remain in the submitted data.

Table 15: S−H nRMSE differences. Negative favors S. Nominal intervals use paired draw resampling conditional on four saved bases and three fitted seeds; finer results use one draw per Re descriptively. Output and seed summaries are nested measurements, not independent samples.

Quantity	Nominal	95% interval	Finer
All 192 effects	+.00232	$[-.00219, +.00496]$	+.00028
Internal 120 effects	+.00221	$[-.00412, +.00580]$	+.00023
Centered response	+.00505	$[-.01791, +.01787]$	+.00100
Shapley allocation	+.00276	$[+.00069, +.00417]$	$-.00035$
All-edge differences by output (averaged over fits) or fit (all outputs):			
Mean drag	+.002114	—	+.005219
Mean lift	+.000140	—	+.000473
Endpoint lift	+.002438	—	+.001943
Rendered near-wake proxy	+.002984	—	$-.006274$
Seed 17	+.005166	—	+.006842
Seed 23	+.000108	—	$-.004431$
Seed 41	+.001681	—	$-.001581$

D.3 Diagnostic-family origin and retrospective CNN example

The full-LOO family arose from an earlier context-MLP diagnosis, before the CNN display. For one previously exposed Re100 parent, that MLP (seed 17, 1,200 updates) predicts mean-drag effects $+.001800420694/-.000640628859$ for action 0 at masks 14/48, versus nominal-reference $+.000676226616/+.000575208664$. Its predicted sign change is absent from that reference; no finer-step check was performed. This is a nominal-reference diagnostic, not prospective CNN-based menu selection.

For the CNN example, we sort the twelve draws by all-edge H−R error and display the seventh (upper median): Re80, third perturbation draw, positive history, seed 17, action 0 and mean drag. Backgrounds 0, 14 and 62 contain zero, three and five other controls. This retrospective rule selects neither independent confirmation nor the largest gain. Its draw-level difference is $-.090533$; the distribution includes the adverse Re60 a2 draw ($+.004321$).

The displayed full-action configuration is in the common training set: Re80, the same saved base, phase category 1 and six coefficients of $.08$. Training reaches its parent through 68 zero-control commands; the test draw undergoes an independent native velocity/pressure perturbation, integration/recovery and a positive $.04$ pulse history. Direct four-frame comparison gives relative L^2 distance $.00671759$ (RMSE $.0108511$), without alignment or spatial weighting. This measures nearby rendered-vorticity inputs with distinct histories, not native-flow energy or far-out-of-distribution generalization.

At masks 62/63, nominal-reference responses are $.00774972/.00808079$ and initial predictions are $.00476021/.00752200$. The initial action-0 difference $.00276180$ is 8.34 times the reference $.00033107$, although the full response is close. These are full responses and conditional effects, not Shapley values. Continuation, R, H and H-plus-loss predict this effect as $.00141027$, $.00174474$, $.00012338$ and $.00005344$; H is closest here. On internal edge $14 \to 15$, reference/initial/R/H effects are $.00021371/.00153601/.00097406/.00041903$.

The H-plus-loss CNN separately illustrates cancellation under background averaging: its action-0 Shapley value $.000201924$ is only 1.81×10^{-6} from reference $.000200109$, but its 32-effect RMSE is $.00020810$; edge $10 \to 11$ is $.00063953$ versus reference $.00016731$. The submitted arrays retain all 192 edges, four outputs, 24 histories, three fitted seeds and six original arms, with null-centering and array keys documented.

D.4 Future-step sensitivity and hypothesis-panel scope

The complete-game numerical check uses the first preassigned draw from each Re base, both histories and fixed model predictions. All $4 \times 2 \times 64 = 512$ endpoints use future step $.005$ (32 reused, 480 new); parent construction, history replay, grid, geometry, waveform and $.05D/U$ command duration

remain unchanged. Tables 16 and 17 compare these same four draws with nominal .01, using the scales and within-draw averaging above. One draw per Re permits descriptive comparisons, not new-source confidence intervals.

Table 16: Complete-game nRMSE on the same four preassigned draws under two future steps. Predictions and training-output scales are fixed; histories/fits are averaged within draw, then draws equally. The grid, parent construction and histories are unchanged. These descriptive means are distinct from the twelve-draw nominal estimates and the older hypothesis panel. Bold/underline rank R, Rφ, H, S and B mean within each error column.

	Nominal .01			Finer .005		
Role	Response	Effect	Shapley	Response	Effect	Shapley
Initial	.420	.280	.231	.583	.291	.253
Ordinary continuation	.380	.231	.187	.556	.248	.215
Matched coverage arms (196 labels)						
R	.36493	.22487	.18570	.54179	.24108	.21116
Rφ	.23925	.15145	.11293	.41048	.17780	.14925
H	**.17522**	**.10509**	**.07111**	**.35653**	**.13129**	<u>.12694</u>
S	<u>.18911</u>	.10901	.07459	<u>.35753</u>	<u>.13156</u>	**.12659**
B mean	.19427	<u>.10575</u>	<u>.07375</u>	.37102	.13746	.12901
Individual random menus and additional feedback controls						
B1	.19129	.10193	.07174	.38057	.13886	.12965
B2	.20797	.11496	.08018	.37760	.14330	.13254
B3	.18355	.10035	.06933	.35489	.13023	.12485
H + full-LOO loss	.173	.104	.071	.353	.129	.126
All consulted information	.147	.071	.053	.314	.106	.107

Every Boolean menu retains lower aggregate effect error than Rφ at .005. B1's ordering against H reverses, B2 remains worse and B3 remains slightly better. B mean is worse than H at both steps on these four draws and becomes worse than S at .005. At the finer step, H is below R in every source/output comparison on both all-edge and internal-edge scores; the information control remains strongest. Nevertheless, all six original roles have higher aggregate effect errors against the finer reference. The reference's own all-edge standardized change is .10827 (Re170: .17848), and Re170's full-game endpoint-lift H$-$R margin narrows from $-.14644$ to $-.05561$.

The preassigned two-edge check uses action 0 at backgrounds 0 and 62, both histories and the first draw per Re (32 branches at .005). Its Re170 endpoint-lift error ordering reverses, prompting eight selected branches at .0025 (Table 17). At that step the difference remains positive for all fits ($+.193456, +.171848, +.153000$), while drag, mean lift and proxy differences remain negative ($-.374230, -.057762, -.109135$). The four physical endpoint-lift effects remain negative at every tested step; their .0025 values are $-.0298244, -.0307884, -.0312723, -.0309635$. Thus the model-error ordering reverses, not the physical sign. Magnitudes have not plateaued. These future-step checks establish neither spatial nor continuum convergence, and do not refine the remaining eight nominal draws.

Separate hypothesis panel. The earlier four-source, eight-history panel uses different normalization and equal condition-level averaging, so it is not pooled with either panel above. Its H model improves response metrics over R but has worse ordinary Re140 selection error ($.0725$ versus $.0654$). Testing all 64 masks under two future steps gives 1,024 branches and 192 action–output event types: one positive, 164 negative and 27 unresolved; all 18 fitted models miss the positive event. Appendix F.3 gives the paired-step rule and nominated-witness results. Response gains therefore do not establish improved discovery yield.

Table 17: Numerical scope of H−R comparisons. Top: standardized effect-error differences across the same four preassigned draws; each finer-step output/edge-family comparison favors H in every draw. Bottom: the selected Re170 two-edge check, averaged over two histories and three fixed fits. Its .0025 step was chosen after inspection; it is neither a new-source interval nor a full-game refined estimate.

	All 192 edges		Internal 120 edges	
Output	Nominal	Finer	Nominal	Finer
Four-output aggregate	−.1198	−.1098	−.1131	−.1119
Mean drag	−.1504	−.1544	−.1437	−.1541
Mean lift	−.0302	−.0275	−.0295	−.0266
Endpoint lift	−.0701	−.0516	−.0652	−.0535
Rendered near-wake proxy	−.1716	−.1537	−.1611	−.1580
Future Δt	Endpoint H	Endpoint R	Endpoint H−R	4-output H−R
.0100	.110477	.292707	−.182230	−.393533
.0050	.343997	.221835	+.122162	−.193205
.0025	.480110	.307342	+.172768	−.081863

E Heterogeneous Conduction: Same-Label Response Learning

This section tests whether singleton-response supervision improves held-out action effects and background dependence after accounting for labels, training budget and initial auxiliary-gradient scale. Longer ordinary training absorbs the original T500 mean advantage: its paired differences from P1500 and P1590 span zero, without establishing equivalence. The evidence therefore includes long-budget T1500 controls, a separate frozen-model modulation confirmation and a new finite-weight comparison with independent material confirmation. The five evaluation cohorts in Table 18 are never pooled; three fitted seeds measure training variation without increasing physical sample size.

Table 18: Distinct conduction panels under the same material distribution and 16-coalition evaluation. "Fresh" refers to materials, not a new distribution.

Material IDs (N)	Fixed models	Evidential role
30000–30015 (16)	B/P500/A500/T500	Previously inspected panel, repeated fits
40000–40063 (64)	Same B/P500/A500/T500	Prespecified fresh-material T500–A500 effect comparison; auxiliary gradients uncalibrated
50000–50063 (64)	T1500/P1590/calibrated A1500, plus schedule controls	Prespecified long-budget effects; later Shapley and modulation analyses secondary/exploratory
60000–60063 (64)	Same frozen long-budget fits	Modulation analysis fixed before generation; all-edge effects secondary
84000–84063 (64)	New T4/A4/P1590 fits	Independent confirmation after separate three-weight selection on IDs 83000–83015

E.1 Physical task and shared-label training

The reference solves $-\nabla\cdot(k\nabla T)=0$ on the unit square, with positive conductivity, left/right temperatures 1/0 and insulating top/bottom boundaries. Each material's smooth log-conductivity uses 24 integer Fourier wave vectors, independent Gaussian sine/cosine coefficients, weights proportional to $(1+p^2+q^2)^{-3/2}$ and ensemble standard deviation .6. Four actions add Gaussians to $\log k$, with peak $\log 2$, width .13 and centers $(.25,.25),(.25,.75),(.75,.25),(.75,.75)$: $k_S = k_0\exp(\sum_{i\in S}\log(2)G_i)$, without material or coalition renormalization. The shared readout is mean temperature over $[.5625,.8125]\times[.375,.625]$. The PDE is linear in temperature at fixed conductivity; actions change its operator.

The 32^2 cell-centered finite-volume solver uses harmonic face conductivities and half-cell Dirichlet distances. Checks cover constant-k and two-layer solutions, residuals, positivity, global-k rescaling and energy–total-flux equality. On two materials and masks 0, 1, 8 and 15, refinement to 64^2 changes

ROI endpoints and null-relative responses by at most 2.12×10^{-4} and 7.27×10^{-5}. These limited checks do not bound every edge's discretization error. Four prespecified materials within eight development games have effect RMS .02078–.02443, background-mean-centered variation 5.33–7.09% of that scale and singleton-additive reconstruction RMSE .001475–.002146; no same-action sign reversal occurs. This pilot's centering over eight backgrounds differs from the singleton-centered modulation below.

A four-level U-Net (Ronneberger et al., 2015) (base width 32; 1,928,417 parameters) maps standardized $\log k$ and two $[-1, 1]$ coordinate channels to the complete temperature field, converted to physical units using training statistics. Boundaries and conservation are not hard-coded. Baseline B uses 128 training materials, 16 uncontrolled validation materials, 1,500 AdamW updates, batch 16, cosine rate $10^{-3} \to 10^{-5}$ and weight decay 10^{-4}. Validation-field MSE selects updates 950/1100/1350 for seeds 17/23/41.

A field-only decision before action inspection expanded the initial seed-17 training set from 64 to 128 materials: controlled-field relative L_2 improved from .019264 to .015574 and effect RMSE from .005764 to .005060, while absolute ROI RMSE worsened from .007942 to .008651. Across those same eight development materials, dividing each material's 32-edge RMSE by its reference-effect RMS and then averaging gives 22.8% for B. Its mean field relative L_2 is 1.56%, with a distinct field-norm denominator. All four initial/final seed-17 checkpoints already predict the 256 eligible development-edge signs correctly under the fixed .001 magnitude screen. The eight diagnostic games overlap validation; the later sixteen-material panel was inspected before the three-fit follow-up.

Within fit, P, A_{abs} and T start from the same B and share 32 training materials' four singleton–null pairs (128 new solves, reused nulls), uncontrolled replay, normalization, sampling and optimizer. Each update uses eight pairs and eight replay fields. P minimizes paired/replay field MSE; A and T add absolute-ROI and response terms, respectively:

$$\mathcal{L}_{\mathrm{resp}} = \frac{[(\hat{q}_i - \hat{q}_0) - (q_i - q_0)]^2}{\sigma_B^2}, \qquad \mathcal{L}_{\mathrm{abs}} = \frac{(\hat{q}_0 - q_0)^2 + (\hat{q}_i - q_i)^2}{2\sigma_B^2}. \tag{16}$$

Here q_0, q_i are reference null/singleton ROI means and σ_B^2 is training-temperature variance. T and A reuse identical endpoint labels; only four singleton edges per training material are directly supervised. The original recipe uses $\lambda_T = \lambda_A = 1$ and fixed update 500, without game-based checkpoint selection. Median gradient norms are $.058/.054/.078$ for A and $.041/.039/.045$ for T across fits, without clipping, motivating the calibration below.

E.2 Long-budget and initial-gradient controls

The disjoint IDs 50000–50063 were generated after fixing the design, using all coalitions once without replacement. Training adds no labels. On each fit's first batch, unit-weight auxiliary-gradient norms g_T, g_A set $\lambda_A = g_T/g_A$ and $\lambda_T = 1$. The finite, nonzero ratios are $.111725/.293010/.224731$ for seeds 17/23/41. This matches initial auxiliary norms, not gradient directions or later dynamics; thirty later diagnostics show changing ratios.

New P, calibrated A and T use cosine schedules with $T_{\max} = 1500$ and fixed checkpoints 500/1000/1500. Their update-500 learning rate, 7.75×10^{-5}, differs from the original 500-update schedule's final 10^{-5}. A separate calibrated A500 retains that original schedule, as do the unchanged original P/A/T500 controls. A 20-update pilot measured .183482 s/P update and .194101 s/T update. The prespecified rule $\max(1500, 10\lceil 1500 t_T/(10 t_P)\rceil)$ assigns an independently fitted P1590 with $T_{\max} = 1590$. Actual P1590 times are 309.0/305.4/307.7 s, versus T1500's 287.0/290.4/290.7 s and calibrated A1500's 288.2/293.9/298.3 s. P therefore receives 5.2–7.7% more wall time; these times include checkpoint/evaluation overhead and exclude simulation.

Raw effect RMSE is computed over 32 edges within material and fit, then averaged over three fits and 64 materials. Whole-material percentile intervals use 5,000 resamples (RNG 20260924), conditional on fixed fits. Table 19 retains every original/long-schedule comparison. T500 beats calibrated A500 by 26.5%, but its differences from longer ordinary fits are unresolved; those fits also have lower null and controlled-field errors on all 64 materials. At the longer budget, T1500 reduces effect RMSE by 15.0% versus P1590 and 13.0% versus calibrated A1500, with every fit mean favoring T.

Controlled-field relative L_2 is .010483 for T1500, .010706 for P1590 and .010983 for calibrated A1500. T−P null-field difference remains unresolved at $-.0000335$ $[-.000127, +.0000653]$. These

Table 19: Long-budget panel (IDs 50000–50063): all-32-edge ROI RMSE and paired contrasts in 10^{-3} temperature units. Means average materials equally within each of three fixed fits. Intervals resample 64 whole materials; only the first two contrasts are prespecified long-budget comparisons. Favorable counts average fits within material.

Recipe	Seed 17	Seed 23	Seed 41	Mean
Original P500	2.999	3.549	4.063	3.537
Original A_{abs}500	2.832	3.408	3.681	3.307
Calibrated A_{abs}500	2.892	3.402	3.839	3.378
Original T500	<u>2.210</u>	2.695	2.545	2.483
P1500	2.292	2.588	2.608	2.496
P1590 cost control	2.268	2.564	2.554	2.462
Calibrated A_{abs}1500	2.223	<u>2.522</u>	<u>2.466</u>	<u>2.404</u>
T1500	**1.984**	**2.324**	**1.968**	**2.092**
Contrast	Difference	Pointwise 95% interval		Favorable
T1500 − P1590	$-.3700$	$[-.4347, -.3077]$		59/64
T1500 − calibrated A_{abs}1500	$-.3115$	$[-.3660, -.2584]$		60/64
Original T500 − calibrated A_{abs}500	$-.8943$	$[-1.0487, -.7356]$		58/64
Original T500 − P1500	$-.0124$	$[-.0662, +.0431]$		35/64
Original T500 − P1590	$+.0213$	$[-.0301, +.0744]$		33/64

comparisons support the response advantage under the tested budget and initial-scale controls; they do not establish equality of optimization dynamics.

Secondary attribution and reference checks. Exact four-action Shapley RMSE is $.0019540871/.0023102438/.0022545999$ for T/P/A. T−P is $-.0003561567$ $[-.0004249243, -.0002886835]$ (15.42%; 57/64 favorable); T−A is $-.0003005128$ $[-.0003579384, -.0002431610]$ (13.33%; 58/64). These post-result, pointwise intervals use the same material resampling and are separate from the next panel. Across 1,024 reference solves, maximum relative residual, flux imbalance and energy–flux discrepancy are 4.57×10^{-15}, 1.01×10^{-13} and 9.53×10^{-14}. Saved scalar games, per-material errors and calibration records accompany the replay.

E.3 Finite weight selection and independent material confirmation

We next repeat the shared-label recipe with new training materials and a symmetric, finite auxiliary-weight comparison. This study is separate from the historical fits above. Training uses IDs 81000–81127, uncontrolled validation 82000–82015, development 83000–83015 and confirmation 84000–84063. The first 32 training materials supply the same 128 singleton–null pairs to all arms. Each seed (17/23/41) shares B, normalization, endpoint labels, the first calibration batch and the first 1,500 pair/replay draws across arms. B trains for 1,500 updates; uncontrolled validation-field MSE selects updates 1150/1350/1350. T and A use 1,500 fixed updates and P uses 1,590.

The prespecified candidates are $\alpha \in \{.25, 1, 4\}$, with $\lambda_T = \alpha$ and $\lambda_A = \alpha g_T / g_A$ in Eq. (5), where g_T, g_A are the unit-weight auxiliary-gradient norms. Calibration reads only the shared first training batch and leaves the sampling RNG unchanged. The three ratios are $.200961728/.297224632/.145235924$; they match initial auxiliary norms, not subsequent gradients. The complete matrix contains 3 B fits, 18 T/A candidates and 3 P fits. For each arm separately, one global α minimizes mean 32-edge RMSE over the same 16 development materials and three fits, with center-first tie order $1, .25, 4$. All candidates are valid and no tie occurs. Both arms select $\alpha = 4$ (Fig. 6); neither is selected separately per seed or by maximizing T−A. The selected configuration and nine checkpoint identities were frozen before confirmation arrays existed. Only selected T4/A4/P fits were then evaluated, without refitting or extending the grid. Both selections lie at its upper boundary, so this is a finite comparison, not a claim to have located an optimal continuous weight.

Confirmation and heterogeneity. Each material/fit contributes a 32-edge RMSE; paired fit differences are averaged within material before averaging the 64 materials. The primary T−A difference

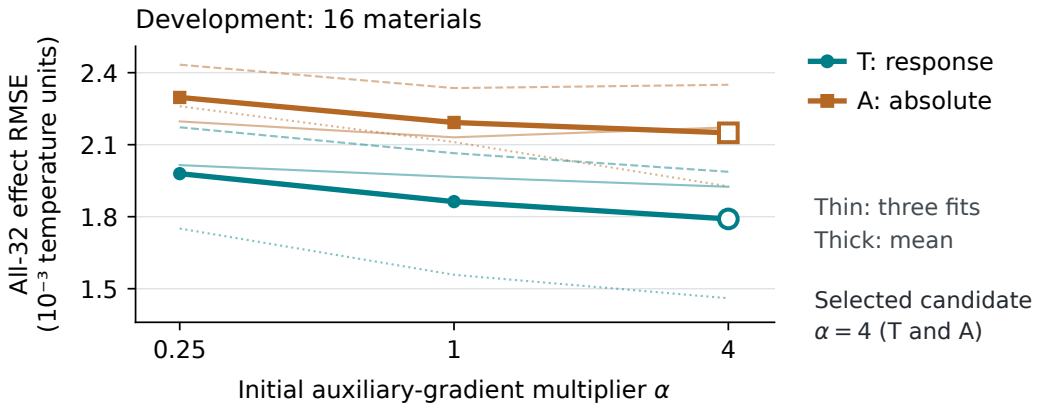


Figure 6: Finite weight selection on 16 development materials. Thin lines show the three fixed training fits; thick lines show means across materials and fits. Open markers indicate $\alpha = 4$, selected for both T and A among $\{.25, 1, 4\}$. Connecting lines do not imply an optimum between or beyond these candidates.

is $-.000437591$ $[-.000497526, -.000379725]$, a 19.4% reduction in the ratio of mean errors (Table 20). Whole-material paired bootstrap intervals use 5,000 resamples (RNG 20260925), conditional on the three fixed fits and development selection. Physical $N = 64$, not 192. All three fit means and 64/64 material-level fit-averaged differences favor T. The secondary T−P difference is $-.000686154$ $[-.000777301, -.000593803]$, with 62 favorable and two adverse materials. The remaining outcomes are secondary, with pointwise descriptive intervals.

Table 20: Independent 64-material confirmation after finite development selection. All entries are RMSE or paired RMSE differences in 10^{-3} temperature units; lower is better. Panel (a) retains each fixed fit. Panel (b) averages fits within material; brackets are 95% whole-material paired intervals conditional on the selected fixed fits. Only all-edge T−A is primary.

	T4	A4	P1590	T−A
(a) All 32 edges, by training seed				
17	**1.9357**	2.3175	2.4228	−.3819
23	**1.8173**	2.3021	2.5604	−.4848
41	**1.6947**	2.1408	2.5230	−.4461
(b) Material means over the three fixed fits				
All 32 edges	**1.8159**	2.2535	2.5021	−.4376 [−.4975, −.3797]
Nonempty-background edges	**1.8214**	2.2625	2.5094	−.4411 [−.5011, −.3831]
Modulation	**.8783**	.9740	.9643	−.0957 [−.1234, −.0678]
Own-singleton replacement	**2.5064**	2.7084	2.8344	−.2020 [−.2553, −.1476]
Shapley	**1.6984**	2.1369	2.3977	−.4385 [−.5018, −.3768]
Absolute ROI	**4.1025**	4.7213	5.0919	−.6188 [−.8080, −.4376]
Null field	**4.9198**	5.3225	4.9951	−.4026 [−.4624, −.3440]
Controlled field	**6.1289**	6.8835	6.3058	−.7546 [−.8185, −.6951]

For the 28 nonempty-background edges, T's actual-effect error is lower than its own-singleton replacement by .000684974 (difference CI $[-.000810214, -.000561671]$), with six adverse materials. Modulation subtracts each model's singleton prediction as defined below; both predictions enter the resulting contrast error. T−A modulation has ten adverse materials despite its favorable mean. T−P modulation is $-.000086043$ $[-.000123736, -.000049236]$, but 19 materials worsen and seed 17 has a positive mean difference $+.000005285$. A−P modulation is $+.000009619$ with an interval spanning zero. A also has higher null/controlled-field errors than P despite lower absolute ROI error. Field entries here are raw temperature RMSE, not relative L_2; controlled-field RMSE pools

equal-sized pixel MSEs over masks 1–15 before taking each material/fit's square root. These findings support the tested response objective without asserting uniform gains or a unique learning mechanism.

Actual costs and replay. The 24 effective fits take 36,270 updates on an NVIDIA GB10. Summed training time is 472.28 s and summed trainer time 537.18 s; the observed first-start to last-finish window is 358.96 s, including checks and handoffs with at most two concurrent fits. Mean selected-fit trainer times are 23.23/23.22/21.06 s for T4/A4/P. P1590 therefore supplies more fixed updates, not a wall-time match on this machine. Maximum CUDA allocator usage is 122.67 MiB. There are 1,552 unique and actual new reference solves: 128 training nulls, 16 validation nulls, 128 singletons, 256 development and 1,024 confirmation endpoints. Thirty-two paired nulls are reused in memory; labels are shared across candidates and seeds. Solver time totals 1.93 s; generation and export take 3.73 s. Engineering smoke costs are separate (four tiny fits, 46 solves). Maximum relative residual, flux imbalance and energy–flux discrepancy are 4.76×10^{-15}, 8.29×10^{-14} and 7.59×10^{-14} at the fixed 32^2 discretization. The accompanying `data/v80_weight_robustness/` stores all candidate development values, selected confirmation scalars, material differences, costs and checkpoint identities. Its `recompute_saved.py` reproduces the reported aggregation and intervals from saved predictions without training; this replay passed on DGX and Windows. Results are not pooled with any earlier material panel.

E.4 Background dependence beyond the supervised singletons

We evaluate how action effects change across backgrounds by subtracting each model's own singleton effect:

$$E_{\mathrm{mod}} = \left[\frac{1}{28} \sum_{i=1}^{4} \sum_{\varnothing \neq S \subseteq N \setminus \{i\}} \left\{ [\hat{d}_i(S) - \hat{d}_i(\varnothing)] - [d_i(S) - d_i(\varnothing)] \right\}^2 \right]^{1/2}. \tag{17}$$

Writing prediction error as $\epsilon_i(S) = \hat{d}_i(S) - d_i(S)$, the contrast error is $\epsilon_i(S) - \epsilon_i(\varnothing)$: both background and singleton predictions enter this measure. The 28 contrasts exclude the four supervised singleton relations and use no evaluation-label correction or action rescaling. Table 21 separates an exploratory application to saved predictions on 50000–50063 from confirmation on 60000–60063. The latter's analysis and T1500/P1590/calibrated A1500 models were frozen before generating new materials, without retraining or replacement. Both use 5,000 whole-material resamples after averaging three fit errors within material; pointwise intervals condition on the fixed fits.

Table 21: Background dependence on two separate 64-material panels, in 10^{-3} temperature units. Errors are computed within material and fit, then averaged equally over three fits and materials. The own-singleton ablation sets $\hat{d}_i(S) = \hat{d}_i(\varnothing)$ for every background. Models are identical across panels; analysis is exploratory on 50000–50063 and frozen before generating 60000–60063.

Material IDs	Model	Modulation	All 32 edges	Actual 28 edges	Ablation
50000–50063	T1500	**1.000808**	—	**2.093020**	**2.546979**
	P1590	1.169846	—	2.464562	2.750965
	Calibrated A_{abs}1500	<u>1.146356</u>	—	<u>2.405500</u>	<u>2.726322</u>
60000–60063	T1500	**.911923**	**2.061664**	**2.065425**	**2.522670**
	P1590	1.080949	2.437922	2.439209	2.754267
	Calibrated A_{abs}1500	<u>1.054279</u>	<u>2.365882</u>	<u>2.367311</u>	<u>2.717979</u>

On the exploratory panel, T's modulation error is lower by 14.45% versus P and 12.70% versus A. Differences in 10^{-3} temperature units are $-.169038$ $[-.199681, -.138928]$ and $-.145548$ $[-.169389, -.122897]$, favoring T on 56/64 and 59/64 materials and in every fit mean. Reference modulation RMS is 2.013299×10^{-3}. The own-singleton ablation removes each model's background dependence; T's actual 28-edge error beats it by $.453960 \times 10^{-3}$ $[.320352, .585263] \times 10^{-3}$, on 52/64 materials. P and A also beat their own ablations.

On the frozen confirmation panel, modulation reductions are 15.64% versus P and 13.50% versus A: differences are $-.169026$ $[-.194814, -.144149]$ and $-.142355$ $[-.162741, -.122634]$ in table

units. Counts are 60/64 and 63/64, with every fit mean favorable. T's actual 28-edge error beats its singleton ablation by $.457245 \times 10^{-3}$ $[.321152, .585609] \times 10^{-3}$ on 53/64 materials; P and A again beat their ablations. All arms therefore use background dependence, while T predicts it more accurately under this comparison.

Secondary all-edge differences on these new materials are $-.376258 \times 10^{-3}$ versus P $[-.460318, -.289386] \times 10^{-3}$ and $-.304217 \times 10^{-3}$ versus A $[-.371745, -.236232] \times 10^{-3}$, with 53/64 and 54/64 favorable materials and every fit mean favorable. Across the 1,024 new solves, maximum residual, flux imbalance and energy–flux discrepancy are 4.31×10^{-15}, 1.22×10^{-13} and 1.18×10^{-13}. No observations from the exploratory panel enter these confirmation intervals.

E.5 Wider-patch development criterion and reproducible scope

One bounded geometry check widened the four Gaussians from .13 to .22 while fixing centers, peak $\log 2$, material distribution, boundaries and ROI. This changes area as well as overlap. Eight existing development materials (20000–20007) supply all 16 coalitions. Relative modulation is the 28-contrast RMS divided by the 32-effect RMS within material, then averaged equally. Before solving, continuation required a greater than 10% mean increase and increases on at least six of eight materials; this is a practical criterion, not a significance test. Table 22 reports the comparison.

Table 22: All eight development materials: relative modulation (28-contrast RMS divided by 32-effect RMS) at 32^2. Width changes from .13 to .22. The predeclared increase criterion fails despite larger absolute modulation.

Material $2000x$	0	1	2	3	4	5	6	7
Width .13	.08961	.10226	.09818	.08027	.09873	.09539	.09850	.09102
Width .22	.08298	.11740	.10310	.08127	.09522	.09490	.09223	.15008

Mean pairwise patch cosine overlap rises .01677→.22149, but relative modulation rises .094245→.102146 (8.38%), with only four of eight increases. Absolute modulation grows 2.11-fold and overall effect scale 1.96-fold. On two preassigned 64^2 materials, relative modulation increases 4.32% on average, on one of two materials; this subset checks the trend, not all-edge convergence. The criterion fails, so no second geometry, model training or new test panel follows. This is a stopped development extension, not a learning failure under stronger background dependence. The 310 new solves and ten reused nulls have maximum residual 3.01×10^{-15}, flux imbalance 9.08×10^{-14} and energy–flux discrepancy 8.49×10^{-14}; conductivity stays positive and temperatures lie between boundary values.

Runnable example and saved-result replay. The core conduction recipe is provided in `code/conduction/README.md`, with configuration-based label generation, B training, P/A/T continuation, checkpoint evaluation and saved-prediction recomputation. It produces new fits from the stated recipe rather than recovering historical checkpoints bit for bit; the other systems retain their saved-result replay interfaces.

The anonymous submission includes the standalone generation–finite-volume–U-Net–evaluation example at `data/v75_conduction/example/README.md`, alongside saved-game replay and checkpoint identities for the reported analyses. Its verified CPU run uses eight training, two validation and four disjoint test materials, 114 solves, 20 baseline updates and 30 P/T updates each. This smaller-network example demonstrates execution, not reproduction of manuscript effect sizes. Commands, configurations, versions and actual outputs remain in the visible package; measured experiment costs are collected in Appendix B.6.

E.6 Earlier fixed-update evidence and its scope

Table 23 retains both the inspected 16-material panel (30000–30015) and the frozen-model 64-material confirmation (40000–40063). The latter's sample, primary T$-A_{\mathrm{abs}}$ contrast and analysis were fixed before generation. Its twelve checkpoints were unchanged; materials are disjoint from

training 10000–10127, validation 20000–20015 and the inspected panel. Both panels use all 16 coalitions and the original equal-500-update, uncalibrated recipe; neither enlarges the other.

Table 23: Earlier fixed-500-update evidence. Top: all-32-edge RMSE in 10^{-3} temperature units. B is the shared starting checkpoint; P, A_{abs} and T use identical labels and updates within fit, with unit-weight, uncalibrated auxiliary losses. The inspected and fresh panels are separate. Counts give materials favoring T over A_{abs}.

Panel	Seed	B	P	A_{abs}	T	Favorable
Inspected 16	17	4.1361	3.0494	2.9189	**2.2014**	15/16
	23	4.6077	3.3884	3.3058	**2.6725**	12/16
	41	7.5502	4.2336	3.7127	**2.6017**	14/16
Fresh 64	17	4.56437	3.41739	3.33347	**2.48502**	59/64
	23	5.65173	3.82240	3.68813	**2.65699**	58/64
	41	6.49146	3.90185	3.51720	**2.41790**	55/64
	Equal-fit	5.56918	3.71388	3.51293	**2.51997**	62/64

Fresh-64 metric	$T - A_{\mathrm{abs}}$: difference [95% interval]	$T - P$: difference [95% interval]
All 32 effects	$-.9930$ $[-1.1463, -.8380]$	-1.1939 $[-1.3832, -.9989]$
28 nonsingleton edges	$-.9918$ $[-1.1466, -.8350]$	-1.1863 $[-1.3768, -.9907]$
Shapley	$-.9879$ $[-1.1544, -.8203]$	-1.1857 $[-1.3881, -.9807]$
Null field	$-.5945$ $[-.7603, -.4473]$	$-.1711$ $[-.2757, -.0736]$
Controlled field	-1.6528 $[-1.8420, -1.4707]$	-1.0281 $[-1.2585, -.8169]$
Absolute ROI	-1.3549 $[-1.7860, -.9390]$	-1.7804 $[-2.3517, -1.2135]$

All lower-block entries are multiplied by 10^3; field rows are dimensionless relative errors. Only $T - A_{\mathrm{abs}}$ all-32-effect RMSE is primary. Other intervals are pointwise, unadjusted and conditional on fixed fits.

On the inspected panel, T−A averages $-.0008206$ (24.8%); T−P averages $-.0010653$ across fits (sample SD .0004952), with 15/16, 14/16 and 15/16 favorable materials. Mean Shapley RMSE improves from A's .003133 to .002338 and controlled-field relative L_2 from .016285 to .014546. All arms already match 512 eligible edge signs per fit. For seed 17, T−P is $-.0008480$ $[-.0011944, -.0005255]$, but one material worsens by .0003457, seven worsen on each field metric, and null-field difference is unresolved at $-.0000379$ $[-.0003233, .0002428]$. Absolute ROI error worsens on 3/7/5 materials across fits.

On fresh64, three paired fit differences are averaged within each material before the same 5,000-resample analysis. The primary T−A difference is $-.0009929607$ $[-.0011463366, -.0008380154]$, a 28.27% reduction. Per-fit reductions are 25.45/27.96/31.26%, with 5/6/9 adverse materials. After averaging fits, IDs 40041 ($+.00066023$) and 40063 ($+.00012266$) remain adverse. The secondary T−P reduction is 32.15%. The 28 nonsingleton-edge outcome excludes directly supervised singleton relations; it is distinct from modulation, which also subtracts each model's singleton prediction. Fourteen materials worsen on absolute ROI and six on null fields despite favorable secondary means.

All 2,048 fresh64 edges per fit exceed the fixed .001 reference-magnitude screen. A gets 2,048/2,047/2,048 directions right and T gets all right; all arms have correct Shapley signs. Gains chiefly concern amplitudes. This magnitude screen is not a numerical-error bound. Across 1,024 solves, maximum residual, flux imbalance and energy–flux discrepancy are 7.88×10^{-15}, 8.39×10^{-14} and 6.58×10^{-14}; float32 storage changes ROI by at most 4.03×10^{-9}. These checks concern the fixed discretization.

Uncontrolled validation and flux remain separate targets. T's validation-field MSE is slightly worse than P's for seed 17 (3.23495×10^{-5} vs. 3.12030×10^{-5}) and seed 23 (4.51794×10^{-5} vs. 4.45970×10^{-5}); seed-17 development left-boundary flux MAE also worsens (.14307 vs. .14037). Thus the positive fixed-update response results do not imply uniform improvement across materials, readouts or longer budgets.

F PHYSICAL HYPOTHESES: SUPPORT, REJECTION AND MISSED EVENTS

An executable hypothesis specifies an action, background and readout that a reference calculation can test. The first two studies freeze predictions before execution; the final study shows how numerical checks reject apparent reversals and reveal a reversal missed by every model. These tasks do not form a pooled discovery-rate estimate.

F.1 PROSPECTIVE SIGN CHANGES UNDER A COMMON AMPLITUDE

The one-dimensional system is periodic transport–Burgers dynamics, $u_t = -.35u_x - \beta u u_x + .04u_{xx} + f(x)$, on $[0, 2\pi)$, with 128 grid points and nominal RK4 step .002. Four time-constant, zero-mean compensated Gaussian sources act for .8 units with $\beta = 0$, followed by a source-free .2-unit forecast with $\beta = .75$. Centers are $\pi + (-1.5, -.5, .5, 1.5)$, narrow/compensating widths .30/.60 and each source's spatial RMS $.8\alpha$, without coalition renormalization. Local energy is $\frac{1}{2}\sum_x w_x u_x^2$, using normalized periodic-Gaussian weights centered at $\pi + .25$ with width .5.

Initial/repaired predictors are mean-preserving residual CNNs (width 32, three blocks, 18,849 parameters). In each of three data repeats, repair starts from the paired initial checkpoint, adds 512 endpoints (1,024 total), and runs 4,800 updates with a unit-weight four-corner velocity/energy contrast objective. Ordinary development endpoint-energy error selects the repaired family independently of confirmation responses. The visible hypothesis-setup package supplies parent construction, numerical implementation, six run configurations, loss code and saved-prediction replay of nominations and reference checks. It does not supply the trained checkpoints.

The 64 confirmation parents use generator seed 96190364. Eight form a separate check set; the other 56 provide 7,168 dependent brackets (four adjacent amplitude intervals times 32 edges). Candidate events change the sign of $d_i(S;\alpha)$ between adjacent common amplitudes in $\{.75, .875, 1, 1.125, 1.25\}$: background and added action scale together. Two frozen rankings each retain at most 16 strict predicted sign changes, one per parent, with deterministic ties. With predicted effects d_-, d_+ and empty-background counterparts $d_{0,-}, d_{0,+}$, the direct score is $M = \min(|d_-|, |d_+|)$; the explanation score is $M[1 + \min(|d_- - d_{0,-}|, |d_+ - d_{0,+}|)/(|d_-| + |d_+| + 10^{-12})]$. Both use the same saved responses, with no extra reference information.

Predictions and rankings precede reference execution. Support requires opposite signs at both amplitudes under nominal and performed finer time steps. The two lists overlap in eight events: all 24 distinct events on 22 parents pass. Initial and repaired models nominate the same sets, so repair has no demonstrated discovery-yield advantage. In the eight-parent check set, all models find 37 nominal sign changes among 1,024 brackets; complete time-step checks cover the first two parents (256 brackets, eight reversals). Only the selected example has a spatial-grid check; these are sign checks, not continuum bounds.

For preselected parent 44, background 10, action 2, the effect changes $+.0157240 \to -.0216235$ as α rises $.875 \to 1$. Both signs persist at $(n, \Delta t) = (128, .002), (128, .001), (256, .001)$. Writing the future field without the added action as u and its change as δ gives

$$d_i(S) = \sum_x w_x u_x \delta_x + \tfrac{1}{2}\sum_x w_x \delta_x^2. \tag{18}$$

Negative overlap outweighs nonnegative increment energy at the higher amplitude (Figure 7); both terms can change as the common amplitude changes. This explains a negative action effect while energy remains nonnegative, without asserting that individually positive actions combine into a negative response.

F.2 PROSPECTIVE CYLINDER FREQUENCY RESPONSE

Four unused perturbations (seeds 43001–43004) share one saved Re100 base and geometry. A four-component wall-motion menu and readout were fixed before nomination; the ordinary stabilized seed-17 primary and eleven descriptive models exported input-only predictions before reference execution, with no further training. This tests mask 2 minus mask 0, not a full new game. Table 24 reports nominal and reduced-time-step results.

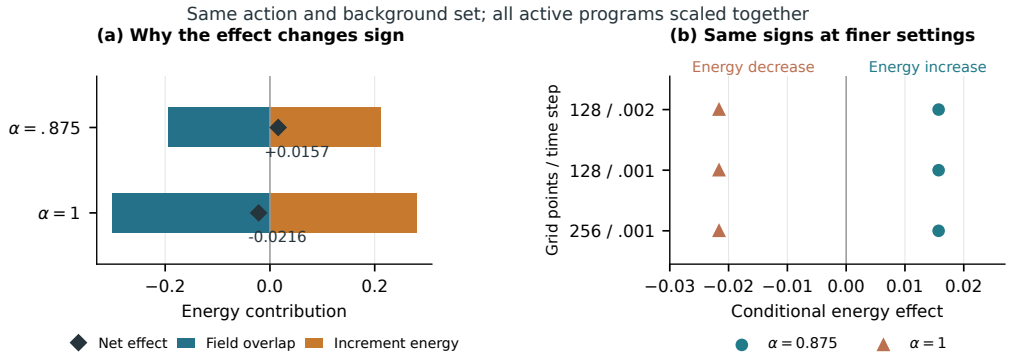


Figure 7: The same action increases or decreases local energy at two common amplitudes. Negative overlap exceeds increment energy at the higher amplitude. Signs persist at the three numerical settings above; spatial refinement applies to this preselected example.

An unforced $60D/U$ scan, discarding its first $10D/U$, fixes $f_0 = .27439955775\,U/D$ (both $25D/U$ halves give about .27432). Components use $(.5, 1, 1.5, 2)f_0$, zero phase and $.08e(t)\sin(2\pi f_j t)$, with $e(t) = \sin^2\{\frac{\pi}{2}\min[1, 2f_0\min(t, 14.6 - t)]\}$ at command midpoints, without renormalization. Mask 2 enables action 1 at f_0. Each branch has 292 control and 146 zero-control commands of $.05D/U$; readouts use the last 146 control commands (7.3–$14.6D/U$): mean command-averaged drag and RMS command-averaged lift. Lift RMS is not necessarily mean-centered. Inputs are four precontrol vorticity frames, action coefficients and fixed context.

Table 24: Frozen primary predictions and all available future-step references for the Re100 frequency test. Effects are mask 2 minus mask 0 in force-coefficient units; each perturbation shares the same saved base flow. A dash is an unrun setting. All directions persist, while amplitudes change.

Perturbation	Readout	Prediction	$\Delta t = .01$	.005	.0025
1	Mean drag	.067586	.067296	.085502	.091777
	Lift RMS	.025247	.025392	.068723	.093541
2	Mean drag	.112593	.110947	.121360	—
	Lift RMS	.186933	.186023	.230655	—
3	Mean drag	.093213	.098019	.111754	—
	Lift RMS	.112040	.122990	.169363	—
4	Mean drag	.112869	.111094	.121464	—
	Lift RMS	.187949	.187161	.231716	—

All eight nominal effects exceed fixed magnitude screens .01 (drag) and .02 (lift RMS). Every one of twelve models, including B, predicts both directions on every perturbation: 96 predictions repeat eight physical relations. Primary nominal MAEs are .002129/.003198. Future-step checks fix history, waveform, command duration, grid and readout; all four perturbations retain positive effects at .005, and perturbation 1 also at .0025. Its lift effect still changes 36.11% on the second halving; at .005, primary MAEs rise to .013455/.047072. Direction persistence is supported, amplitude convergence and a repair advantage are not.

F.3 Rejected reversals and missed ordered witnesses

A historical Re60 parent (domain 1, prefix 63) has full-program drag response/singleton sum $+.002284509/-.002255344$ at future step .01. Same-parent checks give $+.003304625/+.002815604$ at .005 and $+.002285039/+.001881230$ at .0025: the aggregate sign contrast disappears, although nonadditivity remains. Eighteen added branches cover null, singleton and full programs only, with unchanged parents/grid. A different Re60 parent (domain 0, prefix 61) also loses its aggregate sign contrast on refinement.

Reference classification and model nominations. The earlier four-source panel has two histories, six actions and four outputs per source, hence 192 event types. Each asks whether one action has both positive and negative effects across 32 backgrounds. For nominal/halved future-step effects d_n, d_f (.01/.005), define

$$[\ell(S), h(S)] = [d_f(S) - 2|d_f(S) - d_n(S)|,\ d_f(S) + 2|d_f(S) - d_n(S)|]. \tag{19}$$

Parent construction/history remain at .01. This deterministic sensitivity screen is neither a confidence interval nor a certified bound. With $\tau_q = .01 s_q$, where s_q is seed-17 RMS from 168 historical random training responses, an event is positive if $\max_S \ell(S) > \tau_q$ and $\min_S h(S) < -\tau_q$; negative if $\max_S h(S) \le \tau_q$ or $\min_S \ell(S) \ge -\tau_q$; otherwise unresolved. "Negative" means no resolved reversal, not a negative response. Counts are 1 positive, 164 negative and 27 unresolved.

A model nominates its largest/smallest predicted effects when they exceed opposite thresholds, ranked by the smaller absolute effect divided by s_q, with at most six events per history. Ties use action/output order and ascending mask. The saved ordered witness (S_+, S_-) is supported only if $\ell(S_+) > \tau_q$ and $h(S_-) < -\tau_q$; refuted if $h(S_+) \le \tau_q$ or $\ell(S_-) \ge -\tau_q$; otherwise unresolved. A reversal at different backgrounds cannot validate the nomination.

The sole positive event is action 5, Re60 domain-0 negative history, near-wake rendered-enstrophy proxy (output 3). All eighteen arm–fit combinations miss it; none provides a supported ordered witness. Nominal-only evaluation instead gives ten positives and up to two supported witnesses per arm/fit. The two reference-selected witnesses in Table 25 establish the event, not successful model nominations. Response-accuracy gains therefore do not establish discovery-yield superiority.

Table 25: Reference-selected witnesses for the sole positive event, in 10^{-5} rendered-enstrophy-proxy units, with $\tau = 1.2303$. Action 5 adds bit $2^5 = 32$; both transitions belong to one history–action–output event.

Background transition	d_n	d_f	ℓ	h
$31 \to 63$	$+3.3900$	$+3.8877$	$+2.8923$	$+4.8831$
$1 \to 33$	-4.2761	-4.1483	-4.4039	-3.8927

Two prospective smoke nominations predict positive effects while nominal, tight-arithmetic and half-step references are negative. Both fail the fixed screen of three times the across-setting range: two nominations remain unresolved, four abstain and none is supported or refuted. Their saved endpoints are separate from the transfer panels in Appendix G.

G Smoke-flow Feedback: Two Tests and Learning Diagnostics

The smoke-flow tests ask whether action coverage or response losses improve conditional effects when ordinary field fitting does not. Fifty new trajectories test acquisition menus and full-field difference supervision; a separate 24-trajectory panel tests ROI objectives. Neither is pooled with the other or with development data.

G.1 Action menus and paired fields on fifty trajectories

From common factual checkpoints B, random R, structural S and diagnosis-targeted H menus each train ordinary, paired ordinary P and paired full-field difference D models for 400 updates. Each fit uses 32 training trajectories and 1,024 controlled endpoints. Eighteen shared anchors are null, eight singletons, full and eight leave-one-out masks. H adds fourteen internal endpoints forming seven edges for diagnosed actions 1 and 4; R permutes action identities while preserving that graph and coalition sizes; S uses fourteen fixed balanced endpoints. Inputs $[F_4, F_5, F_6, C_7]$ predict full C_8–C_{10} smoke/velocity fields. P supervises a sampled endpoint and its controlled null; D adds their normalized three-field difference with pre-fit gradient-calibrated weight .181083859. Nulls are targets, never inference inputs. P/D share labels, initialization and sampling, but D clips more often, so their optimization dynamics differ.

F continues on all factual four-input/three-future windows from the same trajectories; F_{fixed} repeats $[F_4, F_5, F_6, F_7] \to [F_8, F_9, F_{10}]$ in the controlled menus' 32 sampling slots. Neither adds reference branches. Factual versus controlled comparisons change input diversity and label provenance together. Every continuation in Table 26 is evaluated at fixed update 400.

Each new trajectory supplies 256 coalitions and 1,024 edges. Three fits are averaged within trajectory, then 50 trajectories equally. Effect nRMSE uses each trajectory's reference-effect RMS; controlled-field relative L_2 uses the terminal smoke/two-velocity field. Whole-trajectory bootstrap intervals use 10,000 draws. None of the 400-update arms beats B's mean effect nRMSE. All six prespecified H-minus-R/S acquisition contrasts have Bonferroni 99.1667% intervals spanning zero (Table 27); a targeted-menu advantage is unestablished. Unrounded ordinary R/S/H means are .950774/.950885/.950959: rounded equality does not imply identical predictions, and crossing-zero intervals do not establish equivalence.

Table 26: Fifty-trajectory confirmation (top) and separate four-training-trajectory diagnostics (bottom). Top entries average three fits within each trajectory; bottom rows retain fits separately. Effect/Shapley nRMSE use each trajectory's reference scale; raw effects use ROI-mass units and field relative L_2 the terminal three-channel field. Grouped rows share rounded values only. The training subset has neither intervals nor population claims.

Confirmation arm	Effect nRMSE	Raw effect	Shapley nRMSE	Field rel. L_2
B	**.893**	.00469	**.895**	<u>.338</u>
F	<u>.895</u>	**.00458**	<u>.898</u>	**.337**
F_{fixed}	.951	.00486	.954	.358
R, R_P, S, S_P, H, H_P	.951	.00486	.954	.358
R_D	.917	.00471	.920	.362
S_D	.906	.00468	.909	.362
H_D	.910	<u>.00467</u>	.913	.362
Training-subset fit/arm	**Effect nRMSE**	**Raw effect**		**Field rel. L_2**
17 / B	.958887	.00790868		.332034
17 / Factual fixed-window	1.042557	.00836141		.255975
17 / R	1.044369	.00836110		.255937
17 / H	1.043717	.00835768		.255950
23 / B	1.122481	.00897216		.297260
23 / Factual fixed-window	1.239670	.00950282		.190530
23 / R	1.238568	.00948054		.190434
23 / H	1.238562	.00948849		.190419

Table 27: All six prespecified acquisition contrasts on fifty new trajectories. Signed differences are H minus the comparator, so negative favors H. Whole-trajectory bootstrap intervals use 10,000 common draws after averaging the three fitted seeds within trajectory. The 99.1667% intervals apply the nominal Bonferroni family adjustment across these six comparisons.

Contrast	Difference	Marginal 95% interval	Family-adjusted interval
$H - R$ (ordinary)	$+0.000185$	$[+0.000038, +0.000338]$	$[-0.000006, +0.000398]$
$H - S$ (ordinary)	$+0.000074$	$[-0.000088, +0.000240]$	$[-0.000147, +0.000306]$
$H_P - R_P$	$+0.000095$	$[+0.000019, +0.000174]$	$[-0.000004, +0.000205]$
$H_P - S_P$	$+0.000040$	$[-0.000040, +0.000123]$	$[-0.000070, +0.000157]$
$H_D - R_D$	-0.006603	$[-0.019035, +0.005134]$	$[-0.023709, +0.009177]$
$H_D - S_D$	$+0.003966$	$[-0.007505, +0.015737]$	$[-0.010835, +0.019927]$

D-minus-P effect changes for R/S/H are $-.033800$ $[-.087116, +.015146]$, $-.044422$ $[-.095761, +.003207]$ and $-.040497$ $[-.093944, +.009590]$. Field costs are $+.004430$ $[+.002251, +.006694]$, $+.004607$ $[+.002274, +.006951]$ and $+.004610$ $[+.002364, +.006909]$. These descriptive pointwise 95% intervals lie outside the acquisition family: favorable response means come with worse fields and unresolved response differences.

What training diagnostics establish. Two factual-continuation fits used 256 factual windows from 32 trajectories, batch eight, rate 2×10^{-5} and 400 updates. Mean normalized field MSE in the first/last 50 updates falls .032813→.026869 (seed 17; 18.1%) and .032864→.026287 (seed 23; 20.0%). Different minibatches make these descriptive optimization readings, not paired validation results. Logged gradients are nonzero, with sampled maxima .419/.347.

Table 26 evaluates float64 caches for the first four trajectories of one training shard, a convenience sample. Their unions contain 85/92/86/90 endpoints and 144/153/133/149 edges, including endpoints not acquired by every arm; these are neither complete games nor optimizer losses. Within matched parent/fit, fields improve while effects worsen. Eight separate development trajectories show R/H effect and field errors increasing relative to B; populations remain separate. Controlled–null normalized three-step, three-channel difference RMS values are .00519/.00650/.00255/.00284. These confirm distinct action inputs and targets, not signal-to-noise or convergence; factual-window residuals concern different populations and input distributions. Fitting occurred, but its failure to improve held-out effects remains unexplained.

Factual validation chooses F at update 100 for all fits; response validation chooses R_D400, B and F100 for seeds 17/23/41. On the same fifty trajectories, response-minus-factual selection gives effect nRMSE $+.042557$ $[+.014706, +.072997]$ and field error $+.008313$ $[+.006773, +.009812]$; seed 41 contributes zeros. These are descriptive comparisons, not a targeted-acquisition test. The full fifty-trajectory all-edge reference has not undergone numerical refinement.

G.2 Same-label ROI objectives on twenty-four trajectories

Table 28: ROI confirmation on 24 new trajectories from one shard, averaging three paired fits within trajectory. Top raw effect/Shapley errors use 10^{-3} ROI-mass units; factual ROI errors are unscaled. Bottom raw-effect differences use ROI-mass units. Both intervals in the joint primary comparison (first two rows) cross zero.

Arm	Raw effect	Effect nRMSE	Raw Shapley	Factual ROI	Field rel. L_2
B	6.34	.726	6.33	<u>3.89</u>	.3391
P_{rep}	6.30	.715	6.28	**3.86**	.3432
Q_{absolute}	<u>6.26</u>	<u>.693</u>	<u>6.24</u>	3.99	**.3363**
Q_{response}	**6.07**	**.672**	**6.06**	4.16	<u>.3390</u>
Signed contrast	Mean raw difference		95% trajectory interval		
$Q_{\text{response}} - B$	$-.000273$		$[-.000570, +.000043]$		
$Q_{\text{response}} - P_{\text{rep}}$	$-.000227$		$[-.000482, +.000013]$		
$Q_{\text{response}} - Q_{\text{absolute}}$	$-.000187$		$[-.000452, +.000076]$		

Here q integrates terminal C_{10} smoke over the downstream ROI. Q_{response} supervises $q(S) - q(\varnothing)$ and Q_{absolute} supervises both absolute values, using the same-parent controlled null, complete-field/replay losses, B initialization, 32 training trajectories, pair/replay sequence and 400 updates at three seeds. Each update draws eight pairs and eight factual windows from a seed-specific 32-endpoint structural menu. The common scale is $s_q = .0185118450944$; pre-fit scalar-loss calibration gives weights .0982060781 and $1.4977196557 \times 10^{-5}$, without matching gradients or clipping. AdamW uses rate 2×10^{-5}, weight decay 10^{-5} and norm-one clipping. B and factual-replay P_{rep} are reused. An eight-trajectory development panel favored response accuracy under Q_{response} and factual fields under Q_{absolute}; it is excluded here.

Each confirmation trajectory supplies 256 endpoints and 1,024 edges. Pointwise 95% intervals use 10,000 whole-trajectory resamples after averaging paired fits within trajectory. The joint primary test requires negative upper bounds for both raw-effect contrasts $Q_{\text{response}} - B$ and $Q_{\text{response}} - P_{\text{rep}}$. Neither passes (Table 28); the same-label contrast with Q_{absolute} is also unresolved. Against B, Q_{response} increases factual absolute-ROI error by $+.2648$ $[+.1703, +.3605]$ and nominal direction error by 2.22 percentage points $[+.42, +4.24]$. Per 1,024 edges it corrects 6.76 signs and damages 29.54; five of six early C_8/C_9 factual field-error intervals are adverse. Signs use the nominal reference without a numerical-stability screen.