FLUIDTOPOBENCH: A BENCHMARK FOR LEARNING-BASED FLUID TOPOLOGY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning-based methods are increasingly used in fluid topology optimization to generate channel layouts, predict flow responses, and initialize numerical solvers. However, these capabilities are typically evaluated in isolation, making it difficult to assess how learned outputs translate into physical design quality and downstream optimization performance. We introduce FluidTopoBench, a benchmark of 40,500 steady two-dimensional incompressible Navier–Stokes–Brinkman problems spanning diverse channel layouts, flow regimes, and obstacle configurations. Each instance pairs operating conditions and the response of a uniform initial design with a binary reference topology, its finite-element response and dissipation, and numerical records for downstream evaluation. FluidTopoBench supports topology generation, response prediction, candidate selection, and warm-start optimization under a shared physical evaluation protocol. We evaluate six generators and six predictors using geometric, physical, and downstream metrics across multiple physical and geometric variants. Initial-response conditioning lowers mean dissipation error under unseen boundary layouts, while frozen response predictors provide useful signals for finite-element-verified candidate selection. Generated candidate sets can contain designs with lower dissipation than the numerical reference, and suitable initialization processing improves optimization endpoints. Finally, FluidTopoBench provides a unified framework for evaluating learned geometries and responses by their verified design and optimization outcomes.

1 INTRODUCTION

Topology optimization distributes material to optimize a physical objective under resource and boundary constraints (Bendsøe & Kikuchi, 1988; Sigmund & Maute, 2013). In fluid design, the material layout determines the channels connecting inlets and outlets. We consider minimum-dissipation design for steady incompressible flow, including classical double-pipe and right-angle bend problems, subject to prescribed ports, fixed obstacles, and a fluid-volume budget (Borrvall & Petersson, 2003; Gersborg-Hansen et al., 2005; Alexandersen & Andreasen, 2020). Density-based methods represent fluid and solid regions through Brinkman resistance on a fixed mesh (Alexandersen, 2023). Repeated flow solves make design exploration computationally costly, and the underlying nonconvex optimization is sensitive to initialization. Learned generators, response predictors, and warm starts offer complementary ways to reduce this computational burden (Brunton et al., 2020; Woldseth et al., 2022).

Evaluating these methods requires connecting prediction accuracy to physical design outcomes. Learned channel-generation studies examine geometric reconstruction and conditioning (Guan et al., 2023), while scientific benchmarks measure flow fields, physical quantities, and design rankings (Bonnet et al., 2022; Liu et al., 2024). For fluid topology optimization, a common evaluation must also relate a proposed geometry to its dissipation, a predicted response to candidate selection, and a generated initialization to the optimizer's endpoint. Shared physical instances make these relationships measurable while holding boundary conditions, obstacles, and volume budgets fixed.

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.

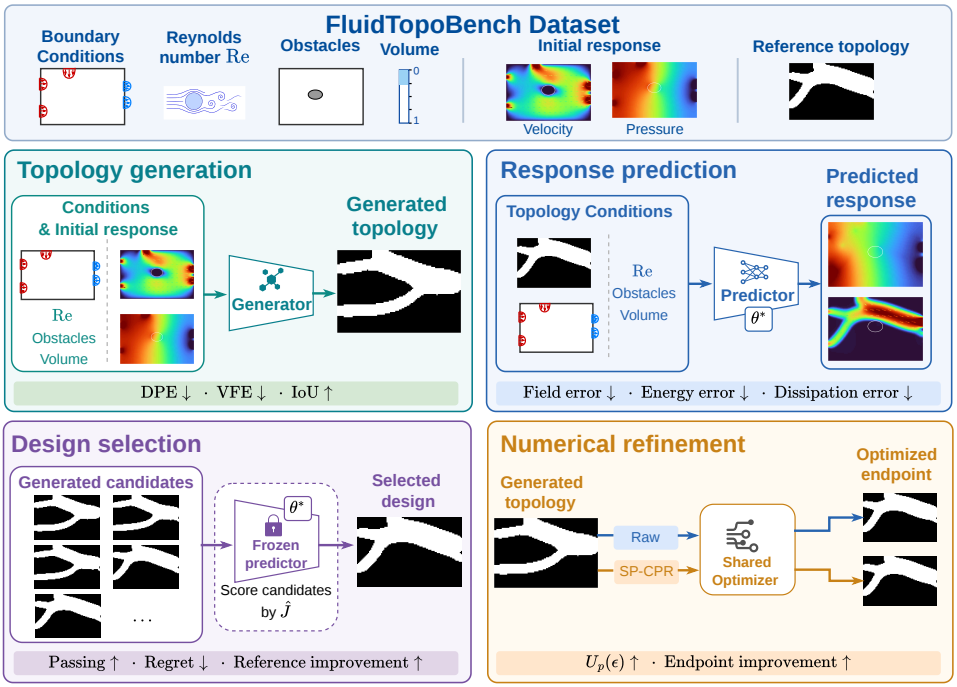


Figure 1: **FluidTopoBench connects learned design, response prediction, candidate selection, and numerical refinement through shared physical records.** Generation uses operating conditions and the uniform-design response; reference-state prediction uses binary geometry and its matched FE target. Frozen predictors score generated candidates before FE verification. Selection and warm-start optimization are separate branches of the same generated output; Raw and SP-CPR pass its floating-point density to the same optimizer.

We introduce FluidTopoBench, a benchmark of 40,500 steady two-dimensional incompressible Navier–Stokes–Brinkman design instances. Each record pairs operating conditions and the response of a uniform initial design with a binary reference topology, its matched finite-element (FE) response and dissipation, and a cold-start optimization record. The dataset spans bends, dual passages, branches, and multi-port layouts, with controlled variations in boundary configurations, flow regimes, and obstacles. These variations test physical generalization. We also evaluate frozen response predictors on generated layouts, using FE solutions on those geometries as targets. The paired records connect topology generation, response prediction, candidate selection, and warm-start optimization (Figure 1).

Our contributions are threefold. First, we provide paired physical data for supervised learning and reanalysis, with controlled design families and held-out conditions for targeted tests of generalization. Second, we define shared evaluation protocols that connect geometric and field accuracy to dissipation, candidate selection, and optimization endpoints. Fixed candidate pools separate candidate availability from selection errors, while paired warm-start experiments compare initialization interfaces applied to the same generated output. Third, six generators and six predictors establish baselines and demonstrate concrete gains: initial-response conditioning lowers mean dissipation error under unseen boundaries, response-guided selection finds FE-verified reference improvements, and initialization processing improves cold-start-parity rates. Controlled edits and response-error comparisons identify location-dependent physical effects and metric-dependent rankings.

2 Related Work

PDEBench (Takamoto et al., 2022) and The Well (Ohana et al., 2024) provide diverse simulation datasets for evaluating learned models of physical systems. RealPDEBench (Hu et al., 2026) combines measured and simulated physical data, while AirfRANS (Bonnet et al., 2022) and FlowBench (Tali

et al., 2025) evaluate flow prediction across varying geometries. For design and inversion, DesignBench (Trabucco et al., 2022) benchmarks offline black-box optimization, AFBench (Liu et al., 2024) evaluates generative airfoil design, and InverseBench (Zheng et al., 2025) benchmarks scientific inverse problems. The proposed FluidTopoBench focuses on fluid material-distribution design, linking layouts to their physical responses and evaluating their use in candidate selection and numerical optimization.

Efficient design exploration must account for both varying physical conditions and the performance of proposed layouts. Generative models can propose topologies without repeating full numerical optimization (Mazé & Ahmed, 2023); initial-domain physical fields encode the spatial effects of boundary conditions (Nie et al., 2021); and numerical refinement can further improve generated designs under the physical objective (Oh et al., 2019). Generative exploration coupled with high-fidelity evaluation also supports thermal-fluid design (Yaji et al., 2022). For fluid channels, Guan et al. (2023) examine field-conditioned topology generation in Stokes flow with geometric and volume objectives. Neural operators approximate physical-response mappings (Kovachki et al., 2023), while active data collection addresses the geometry shifts encountered during optimization (Kou et al., 2025). Goal-oriented error estimation targets errors in physical quantities of interest (Becker & Rannacher, 2001), and decision-focused learning connects prediction quality to downstream decisions (Donti et al., 2017; Elmachtoub & Grigas, 2022). These considerations motivate evaluating generated layouts, predicted responses, candidate choices, and optimization endpoints on shared physical instances.

3 FLUIDTOPOBENCH

3.1 PROBLEM FORMULATION

Let $\Omega \subset \mathbb{R}^2$ be the design domain and $\rho : \Omega \to [0,1]$ the dimensionless design density, with $\rho = 1$ representing fluid and $\rho = 0$ solid. Operating conditions c specify the boundary ports, Reynolds number $\mathrm{Re} > 0$, fluid-volume fraction limit $V_f \in (0,1]$, and fixed obstacles $\mathcal{O} \subset \Omega$. The velocity and pressure fields are $\mathbf{u} : \Omega \to \mathbb{R}^2$ and $p : \Omega \to \mathbb{R}$, respectively. With symmetric strain rate $\boldsymbol{\varepsilon}(\mathbf{u}) = (\nabla\mathbf{u} + (\nabla\mathbf{u})^T)/2$, the minimum-dissipation design problem is

$$\min_{\rho,\mathbf{u},p} \quad \phi(\rho,\mathbf{u}) = \int_\Omega \left[\mu\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u}) + \tfrac{1}{2}\alpha(\rho)|\mathbf{u}|^2\right] \mathrm{d}\mathbf{x}, \tag{1}$$

$$\text{subject to} \quad \rho_f(\mathbf{u}\cdot\nabla)\mathbf{u} - \nabla\cdot(2\mu\boldsymbol{\varepsilon}(\mathbf{u})) + \nabla p + \alpha(\rho)\mathbf{u} = 0, \quad \nabla\cdot\mathbf{u} = 0, \tag{2}$$

$$0 \le \rho \le 1, \quad \rho|_{\mathcal{O}} = 0, \quad V(\rho) := \frac{1}{|\Omega|}\int_\Omega \rho\,\mathrm{d}\mathbf{x} \le V_f. \tag{3}$$

Here $\rho_f > 0$ is the fluid mass density, $\mu > 0$ is the dynamic viscosity, and $\alpha : [0,1] \to [0,\infty)$ is the Brinkman resistance, with low resistance in fluid and high resistance in solid. The colon denotes the Frobenius inner product, so that $\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u}) = \|\boldsymbol{\varepsilon}(\mathbf{u})\|_F^2$. The state equations hold in Ω, with parabolic inlet velocity profiles, prescribed outlet conditions, and no-slip domain walls. The Reynolds number sets the viscosity through the characteristic inlet speed and width; boundary conditions and material parameters are specified in Appendix A.1. For a given geometry, we write $J_c(\rho) = \phi(\rho, \mathbf{u}_c(\rho))$ for its FE-evaluated objective, where $\mathbf{u}_c(\rho)$ is the velocity solution under conditions c.

3.2 DATASET GENERATION AND SPLITS

Reference designs are generated with a TopFlow-based optimizer (Alexandersen, 2023), combining stabilized FE solves with optimality-criteria density updates under the design constraints. Thresholding the uniform-start endpoint at 0.5 gives ρ_{ref}; FE reanalysis supplies its matched response and $J_{\mathrm{ref}} = J_c(\rho_{\mathrm{ref}})$. Records also include conditions c, the uniform-design response, and solver metadata, and generated candidates use the same FE analysis. Multi-start, mesh-refinement, independent-FE, and state-consistency checks support the references as competitive designs with numerically stable physical targets (Appendix A.4).

Boundary templates span dual passages, bends, T/Y-junctions, and multi-port layouts, with perturbed ports and varied obstacle shapes, sizes, and positions. Train, Validation, and Test share the same

sampling distribution and 30-template boundary catalog, with Re sampled logarithmically over $[10^{-3}, 200]$ and $V_f \in \{0.25, 0.30, \ldots, 0.45\}$. Separate out-of-distribution (OOD) splits test unseen boundary templates, a lower-Reynolds regime, altered obstacle configurations, and a joint boundary–resolution shift (Table 1). Construction and sampling details appear in Appendices A.2–A.3.

Table 1: **Dataset composition and evaluation shifts.** In-distribution splits share a 30-template boundary catalog. Grids are 64×96 unless stated otherwise.

Split	Instances	Shift	Conditions
In-distribution splits			
Train	32,000	—	30 boundary templates
Validation	4,000	—	Same distribution as Train
Test	2,000	—	Independent test instances
Out-of-distribution splits			
OOD-BC	1,000	Boundary	10 held-out templates
Low-Re	250	Physics	$\mathrm{Re} \in [10^{-5}, 5 \times 10^{-4}]$
Dual	150	Geometry	Two obstacles
Large	200	Geometry	Enlarged obstacle
Notch	150	Geometry	Boundary-attached notch
HR-OOD-BC	750	Joint	Held-out templates, 128×192

3.3 Benchmark Tasks and Evaluation

Topology generation. Generators map c and the uniform-design response to a density field. We evaluate its binary geometry ρ by FE analysis and compare it with ρ_{ref} using dissipated power error (DPE), volume fraction error (VFE), and intersection-over-union (IoU):

$$\mathrm{DPE} = \frac{J_c(\rho) - J_{\mathrm{ref}}}{|J_{\mathrm{ref}}|}, \qquad \mathrm{VFE} = \frac{|V(\rho) - V_f|}{V_f}, \qquad \mathrm{IoU} = \frac{|\rho \cap \rho_{\mathrm{ref}}|}{|\rho \cup \rho_{\mathrm{ref}}|}. \tag{4}$$

Negative DPE indicates lower dissipation than the reference. These metrics measure objective quality, volume-budget deviation, and geometric overlap.

Response prediction. Predicted velocity $\widehat{\mathbf{u}}$ and pressure $\widehat{p}$ are compared with the FE solution on the input geometry. Alongside relative L_2 errors $E_y = \|\widehat{y} - y\|_2 / \|y\|_2$, $y \in \{\mathbf{u}, p\}$, we report energy-norm error η_ρ and objective error E_J using the dissipation bilinear form:

$$\begin{aligned} a_{\rho,h}(\mathbf{v}, \mathbf{w}) &= \int_\Omega \left[2\mu \boldsymbol{\varepsilon}(\Pi_h \mathbf{v}) : \boldsymbol{\varepsilon}(\Pi_h \mathbf{w}) + \alpha(\rho)(\Pi_h \mathbf{v}) \cdot (\Pi_h \mathbf{w}) \right] \mathrm{d}\mathbf{x}, \\ \eta_\rho &= \sqrt{\frac{a_{\rho,h}(\widehat{\mathbf{u}} - \mathbf{u}, \widehat{\mathbf{u}} - \mathbf{u})}{a_{\rho,h}(\mathbf{u}, \mathbf{u})}}, \qquad E_J = \frac{|\widehat{J} - J_c(\rho)|}{J_c(\rho)}. \end{aligned} \tag{5}$$

Here Π_h is the shared cell-to-node interpolation and $\widehat{J} = \frac{1}{2} a_{\rho,h}(\widehat{\mathbf{u}}, \widehat{\mathbf{u}})$. Reference and generated inputs each use their own FE target and the same frozen predictor. Appendix C.1 specifies interpolation, pressure gauges, and the relation between errors.

Candidate selection. A fixed candidate pool is ranked by volume-based rules or predicted dissipation $\widehat{J}$, without candidate FE labels. The chosen design is FE-verified. We report quality-band passing, reference improvement, and conditional regret to the best FE-valid, volume-eligible candidate. Generation, prediction, and FE-query budgets are tracked separately.

Numerical initialization. Continuous generator outputs $\widehat{\rho}$ initialize the optimizer directly (*Raw*) or after Structure-Preserving Cleanup, Projection, and Relaxation (*SP-CPR*; Appendix H.1). Interfaces share the generated input, optimizer, and stopping rule. Among N attempts under interface P, let $\mathcal{S}_P$ contain completed runs and $\phi_{P,\mathrm{end},i}$ and $\phi_{\mathrm{cold},i}$ denote their warm and matched continuous cold objectives. We report the relative objective excess and joint completion–quality rate:

$$\Delta\phi_{P,i} = \frac{\phi_{P,\mathrm{end},i} - \phi_{\mathrm{cold},i}}{|\phi_{\mathrm{cold},i}|}, \qquad U_P(\epsilon) = \frac{\left|\{i \in \mathcal{S}_P : \Delta\phi_{P,i} \le \epsilon\}\right|}{N}. \tag{6}$$

Failures remain in the denominator; negative ϵ requires improvement over the cold endpoint. Paired improvements and solver iterations are reported separately.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We compare deterministic, adversarial, diffusion, and flow-based generators to assess how generation strategies affect physical design quality. CNN and cGAN (Isola et al., 2017) use U-Net backbones (Ronneberger et al., 2015). DDPM (Ho et al., 2020) predicts diffusion noise; Rectified Flow (RF) (Liu et al., 2023) learns a noise-to-topology flow; and EDM (Karras et al., 2022) uses preconditioned denoising. PhysToken is a DDPM-based conditioning variant implemented for this benchmark, adding physical-condition tokens and scalar feature modulation (Appendix B.1).

Response baselines compare spatial representations and physics-informed training for both field reconstruction and dissipation estimation. U-Net (Ronneberger et al., 2015) provides a multiscale convolutional encoder–decoder; FNO (Li et al., 2021) models global interactions in Fourier space; and U-FNO (Wen et al., 2022) combines spectral layers with local U-Net paths. Our PINO implementation (Li et al., 2024) adds physics-residual penalties to the U-FNO backbone. CNO (Raonić et al., 2023) uses multiscale convolutional operators, while DeepONet (Lu et al., 2021) combines an input-encoding branch with a coordinate-encoding trunk.

Models use validation-selected checkpoints; predictors train on binary-reference FE responses. Generator samples are averaged within instances, with paired instance-bootstrap intervals. Selection macro rates weight generator–split conditions equally; initialization rates retain all attempts. Appendices B and C.2 give architecture, training, and statistical details.

4.2 TOPOLOGY GENERATION AND PHYSICAL GENERALIZATION

Baseline performance. On held-out boundary layouts, every generator has higher mean DPE and lower IoU than on Test (Table 2). RF leads mean DPE and VFE; CNN leads IoU. Equal-budget edits further examine this metric distinction (Appendix E). Figure 2 shows shared-condition layouts; Table 12 reports all splits.

Table 2: **Topology-generation baselines.** DPE and VFE are percentages; IoU is a ratio. Samples are averaged within each instance. Bold marks the best point estimate per column.

Generator	Test				OOD-BC			
	DPE (%) ↓		IoU ↑	VFE (%) ↓	DPE (%) ↓		IoU ↑	VFE (%) ↓
	Mean	Median			Mean	Median		
CNN	0.72	0.09	**0.961**	1.35	2.23	0.43	**0.922**	1.55
cGAN	2.27	0.34	0.945	1.81	3.26	0.79	0.910	2.47
DDPM	8.21	4.33	0.899	4.14	9.33	3.10	0.887	4.02
RF	**0.18**	**-0.22**	0.946	**1.07**	**1.77**	**0.02**	0.921	**1.13**
EDM	2.75	0.09	0.864	2.39	7.05	1.02	0.856	2.65
PhysToken	2.22	0.63	0.929	1.53	3.51	1.00	0.906	1.58

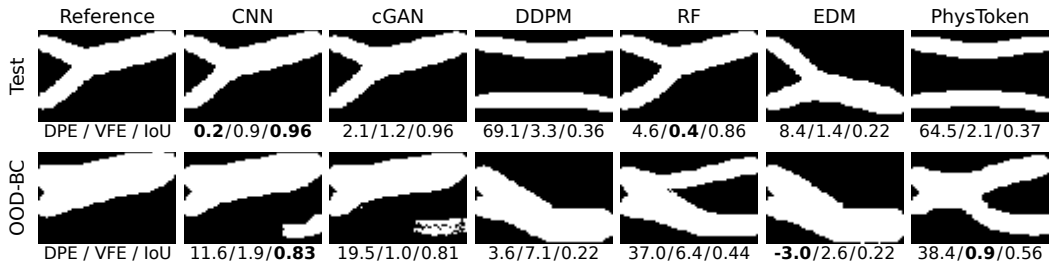


Figure 2: **Generated topologies under shared Test and OOD-BC conditions.** Each row compares the reference with six generators. Labels report DPE (%), VFE (%), and IoU; bold marks the best value per metric within each row.

Initial-response conditioning. Replacing boundary-map conditioning with the response of a uniform initial design lowers mean DPE for all six generators on OOD-BC, while Test effects vary (Figure 3). The two conditioning representations differ in both response information and boundary encoding. Paired effects for all splits appear in Appendix D.1.

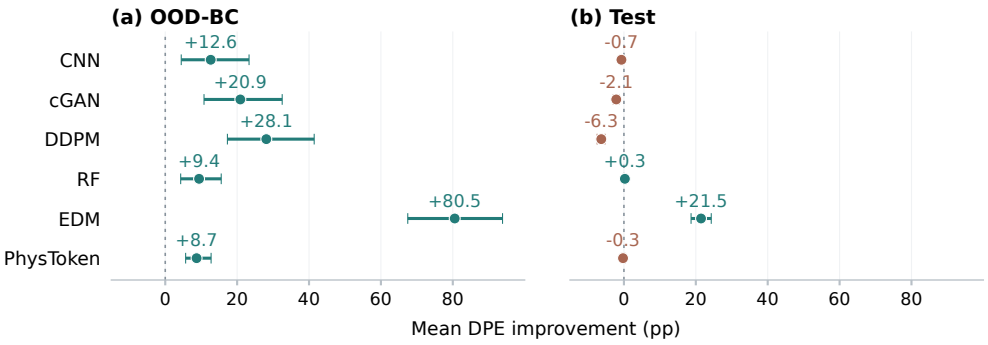


Figure 3: **Initial-response conditioning under boundary shift.** Points show paired mean-DPE differences, boundary-map minus initial-response, in percentage points. Bars are pointwise 95% instance-bootstrap intervals; positive values favor initial-response conditioning.

Equal geometric errors, different physical consequences. Controlled topology edits are constructed with the same size, fluid volume, and geometric overlap, with additional matching of their local geometric context. Despite these controls, their dissipation changes remain sensitive to where and how the edits are introduced, including cases where the relative effect reverses across edit arrangements. The response is also spatially non-local: across the controlled settings, 38.2–63.0% of the absolute dissipation redistribution occurs outside the immediate four-cell neighborhood of the edited region. Thus, similar geometric errors can produce different and non-local physical responses, motivating FE-based evaluation beyond geometric overlap alone (Appendix E).

4.3 Response Prediction on Reference and Generated Designs

Metric-dependent response accuracy. Predictor ordering depends on both the physical setting and the quantity being evaluated (Table 3). On Test, U-FNO and PINO have similarly low velocity errors, U-FNO has the lowest η_ρ, and CNO has the lowest E_J; CNO leads all four OOD-BC metrics by point estimate. These metrics serve different uses: E_u measures global velocity reconstruction, η_ρ emphasizes errors through strain and porous dissipation, and E_J measures the resulting mismatch in the integrated objective.

Table 3: **Response prediction on reference geometries.** Mean dimensionless errors; lower is better. Bold marks the best point estimate before rounding.

	Test				OOD-BC			
Predictor	E_p	E_u	η_ρ	E_J	E_p	E_u	η_ρ	E_J
U-Net	0.079	0.035	0.177	0.080	0.704	0.407	0.561	0.322
FNO	0.054	0.025	0.099	0.041	0.478	0.333	0.433	0.270
U-FNO	0.037	0.014	**0.041**	0.042	0.314	0.124	0.204	0.177
PINO	**0.037**	**0.014**	0.042	0.043	0.321	0.129	0.213	0.179
CNO	0.040	0.016	0.053	**0.040**	**0.245**	**0.071**	**0.125**	**0.102**
DeepONet	0.072	0.080	0.654	0.552	0.546	0.630	4.506	28.427

Figure 4 highlights how the evaluation target affects predictor choice. Velocity residuals and dissipation-weighted errors emphasize different regions of the same predicted flow. Across all standard splits, rankings under E_u and η_ρ agree more closely with each other than with E_J. Dissipation accuracy therefore provides a distinct criterion for choosing response models that will be used to score designs.

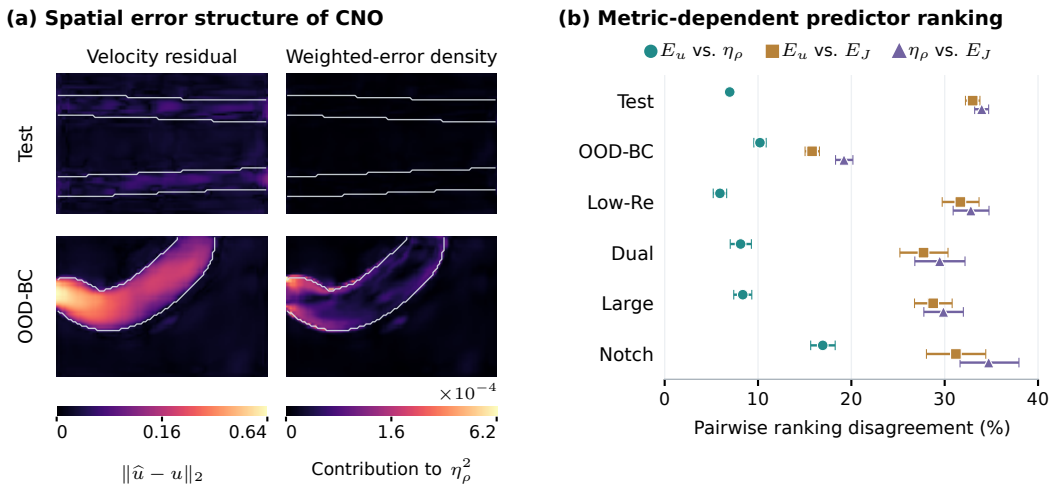


Figure 4: **Response metrics induce different views of predictor accuracy.** **(a)** CNO velocity residual and local dissipation-weighted error on illustrative reference geometries; each quantity shares a scale across examples. **(b)** Within-instance ranking disagreement across E_u, η_ρ, and E_J; bars show pointwise 95% instance-bootstrap intervals. Appendix F.1 defines the comparison.

The dissipation structure helps explain this metric dependence. For the shared quadratic, the objective change combines a weighted squared residual with a cross term involving the target velocity. These contributions can reinforce or partially offset each other, allowing similar η_ρ to yield different dissipation estimates (Appendix C.1). Field reconstruction and objective estimation thus assess complementary properties of the predicted response.

Prediction on generated geometries. Candidate selection requires response models to estimate dissipation on layouts supplied by generators. We assess this transfer with frozen checkpoints and an FE target computed for each input geometry. The paired evaluation shows higher macro mean velocity error for all six predictors and higher weighted and objective errors for five (Table 16(a)). This sensitivity to generated inputs makes the candidate distribution part of evaluating a predictor for design use. The fixed-pool experiments in Section 4.4 then assess how these predictions translate into candidate choices. Appendix F.2 gives the paired evaluation protocol.

4.4 Response-Guided Candidate Selection

Fixed-pool selection. Volume screening identifies eligible candidates; predicted dissipation helps distinguish their physical quality. All selectors receive the same five-candidate pools. Frozen predictors rank without candidate FE labels, and the chosen design is then FE-verified. Under the default band, VFE $\leq 3\%$ and DPE $\leq 10\%$, CNO improves macro passing by 2.50 percentage points over maximum volume and more than halves mean conditional regret (Figure 5(b)). All six predictors improve macro passing (Table 17), and CNO's gain remains positive across nine tested quality bands (Appendix G.1).

Same-pool FE screening evaluates every eligible candidate and establishes the passing rate attainable from the available designs. CNO recovers most of this potential gain over maximum volume. Most remaining CNO failures occur in pools without a passing design, identifying candidate supply as the main remaining limitation on default-band passing. The example in Figure 5(c) illustrates the value of dissipation-based ranking among candidates that all meet the volume criterion.

Reference improvement. We additionally evaluate improvements over the binary reference under the original one-sided volume limit and fixed-obstacle constraints:

$$\mathcal{R}_\gamma(c) = \{\rho : V(\rho) \leq V_f,\ \rho|_{\mathcal{O}} = 0,\ \text{FE valid},\ J_c(\rho) \leq (1-\gamma)J_{\text{ref}}\}. \tag{7}$$

Positive γ specifies the required fractional reduction in dissipation. CNO and maximum volume have similar success rates at reference parity; CNO achieves higher success at every tested positive

improvement margin (Figure 5(a)). Its ranking signal therefore helps recover reference-improving designs within the same proposal budget.

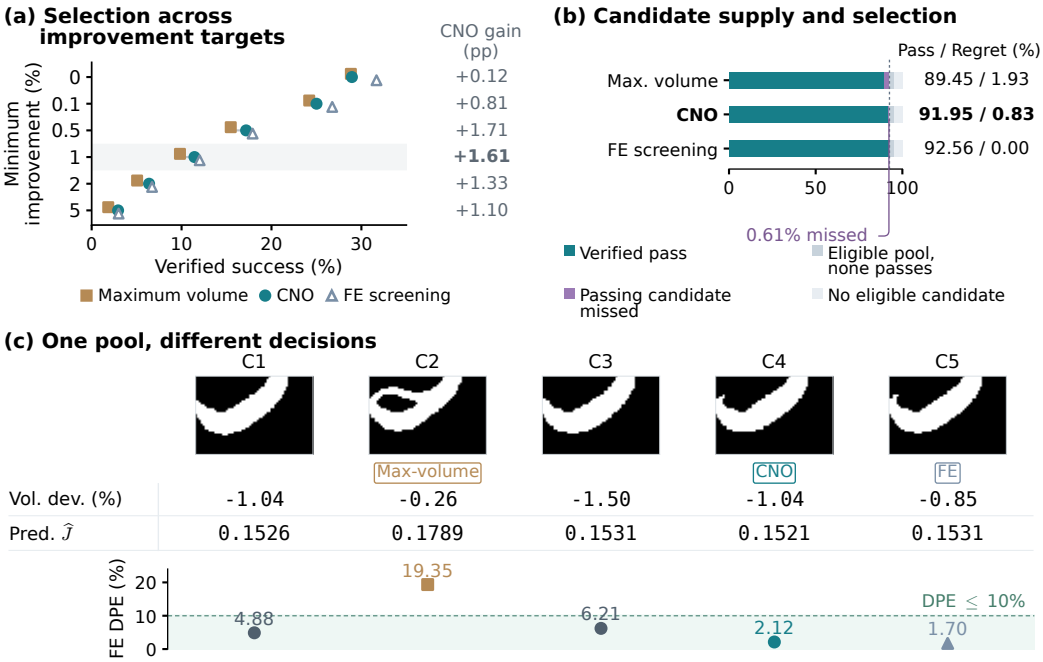


Figure 5: **Response-guided selection from shared candidate pools.** (a) Reference improvement under the volume cap and obstacle constraint. (b) Default-band outcomes and conditional regret. (c) Volume-eligible candidates with predicted scores, FE-verified DPE, and different selections. Aggregate passing rates weight generator–split conditions equally and include all planned pools; FE screening analyzes every eligible candidate.

Generator–predictor collaboration. We combine sixteen RF proposals per instance with CNO selection under the original volume and obstacle constraints. In the six-split macro average, selected designs match or improve on the binary reference in 66.86% of cases, and 28.66% improve by at least 1%. Reference improvements occur across all six splits, demonstrating the combined ability to propose and identify better layouts under the same physical conditions. These rates retain all planned instances, including failed verification. Appendix G.2 reports split-wise results and verified examples.

4.5 Numerical Optimization from Generated Initializations

Initialization interfaces. Generated densities also provide warm starts for numerical optimization. We compare Raw, Projection-only, Global-relaxation, and SP-CPR using the same generated inputs, optimizer, and stopping rule, with endpoint quality measured against the matched continuous cold-start objective. Projection-only enforces the boundary convention and volume budget; Global-relaxation adds mixing toward a uniform design, with its weight selected using the relaxed validation target. SP-CPR applies structure-aware cleanup and local relaxation before projection (Appendix H.1).

SP-CPR favors stricter endpoint quality. SP-CPR achieves the highest joint success rates at the reported improvement targets and at cold-start parity (Figure 6(a,b)). When at least 0.1% improvement over the cold endpoint is required, its rate is 32.1%, compared with 26.9% for Global-relaxation. This advantage holds across all six splits, with paired pointwise 95% intervals above zero (Figure 6(c)). Allowing 0.1% objective excess brings both rates above 92%, with Global-relaxation ahead. Global mixing therefore captures much of the benefit under the relaxed criterion, while SP-CPR more often reaches lower-dissipation endpoints under stricter targets.

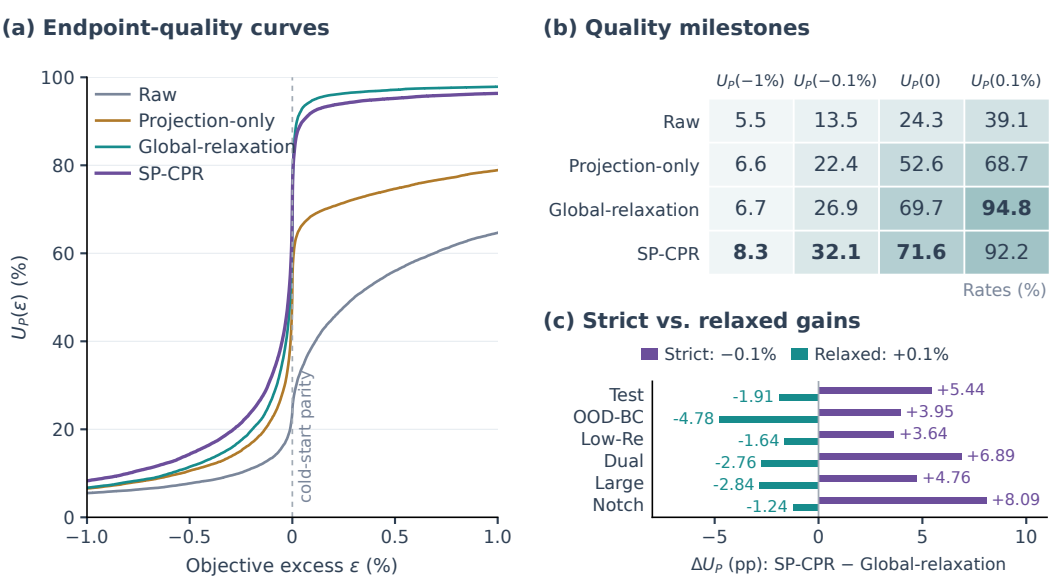


Figure 6: **SP-CPR favors stricter endpoint quality.** (a) Joint completion–quality curves relative to the continuous cold endpoint; negative thresholds require improvement. (b) Rates at selected thresholds. (c) SP-CPR minus Global-relaxation across splits. All rates include every paired attempt, counting failures as non-passing.

Among runs completed under both Raw and SP-CPR, SP-CPR reduces the objective by a median of 0.327 percentage points of the matched cold objective (Appendix H.3). Together, the joint success rates and paired objective reductions show that initialization processing can improve the designs reached from generated densities under a fixed optimization policy.

5 Discussion and Conclusion

FluidTopoBench provides an integrated evaluation of learning-based fluid topology optimization across topology generation, response prediction, candidate selection, and numerical initialization. Using paired physical instances and a shared FE protocol, it connects model outputs to verified design and optimization outcomes. The experiments show that initial-response conditioning improves generation under unseen boundary layouts, response predictors provide useful signals for selecting generated designs, and initialization processing changes the endpoints reached from generated warm starts. Metric-dependent predictor rankings further show that geometric, field-level, objective, and downstream measures capture complementary aspects of performance.

The paired evaluation separates candidate availability, selection quality, and optimizer initialization as distinct sources of downstream performance. Multi-start optimization, mesh refinement, and independent FE checks support the released references as competitive and numerically stable baselines for most assessed problems. Within the studied steady two-dimensional Navier–Stokes–Brinkman setting, FluidTopoBench provides a common basis for developing generators, response models, and numerical interfaces evaluated through FE-verified design outcomes.

REPRODUCIBILITY STATEMENT

Anonymous source code, configurations, evaluation scripts, numerical result records, and a representative paired benchmark subset are available at `https://anonymous.4open.science/r/FluidTopoBench-120C/`. Appendices A–C document the physical construction, model configurations, and evaluation and statistical protocols. A self-contained supplementary archive is also provided with the submission. The complete 40,500-instance dataset and inference checkpoints will be released upon de-anonymization.

ETHICS STATEMENT

All benchmark data are generated numerically and contain no human-subject or personal data. The benchmark studies idealized fluid designs under specified numerical and constitutive assumptions. Manufacturing, safety, and application-specific validation are outside its scope and require additional engineering analysis.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and readability of the manuscript, provide feedback on parts of the experimental design, and assist in creating scientific figures and illustrations. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content.

REFERENCES

Joe Alexandersen. A detailed introduction to density-based topology optimisation of fluid flow problems with implementation in MATLAB. *Structural and Multidisciplinary Optimization*, 66(1): 12, 2023. doi: 10.1007/s00158-022-03420-9.

Joe Alexandersen and Casper Schousboe Andreasen. A review of topology optimisation for fluid-based problems. *Fluids*, 5(1):29, March 2020. doi: 10.3390/fluids5010029.

Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001. doi: 10.1017/S0962492901000010.

Martin P. Bendsøe and Noboru Kikuchi. Generating optimal topologies in structural design using a homogenization method. *Computer Methods in Applied Mechanics and Engineering*, 71(2): 197–224, 1988. doi: 10.1016/0045-7825(88)90086-2.

Florent Bonnet, Jocelyn Mazari, Paola Cinnella, and Patrick Gallinari. AirfRANS: High fidelity computational fluid dynamics dataset for approximating Reynolds-averaged Navier–Stokes solutions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23463–23478, 2022.

Thomas Borrvall and Joakim Petersson. Topology optimization of fluids in Stokes flow. *International Journal for Numerical Methods in Fluids*, 41(1):77–107, January 2003. doi: 10.1002/fld.426.

Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52:477–508, 2020. doi: 10.1146/annurev-fluid-010719-060214.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://arxiv.org/abs/1703.04529`.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Allan Gersborg-Hansen, Ole Sigmund, and Robert B. Haber. Topology optimization of channel flow problems. *Structural and Multidisciplinary Optimization*, 30(3):181–192, 2005. doi: 10.1007/s00158-004-0508-7.

Ce Guan, Jianyu Zhang, Zhen Li, and Yongbo Deng. Data-driven topology optimization of channel flow problems, 2023. URL https://arxiv.org/abs/2309.00278.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. RealPDEBench: A benchmark for complex physical systems with real-world data. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/aed13d8a5ee83a9c6880285b69d0ba81-Abstract-Conference.html.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976. IEEE, July 2017. doi: 10.1109/cvpr.2017.632.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Chenhui Kou, Yuhui Yin, Min Zhu, Shengkun Jia, Yiqing Luo, Xigang Yuan, and Lu Lu. Efficient neural topology optimization via active learning for enhancing turbulent mass transfer in fluid channels, 2025. URL https://arxiv.org/abs/2503.03997.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.08895.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM / IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506. URL https://dl.acm.org/doi/10.1145/3648506.

Jian Liu, Jianyu Wu, Hairun Xie, Guoqing Zhang, Jing Wang, Wei Liu, Wanli Ouyang, Junjun Jiang, Xianming Liu, Shixiang Tang, and Miao Zhang. AFBench: A large-scale benchmark for airfoil design. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/96bc329285d3f028dc5972d7146c7e2f-Abstract-Datasets_and_Benchmarks_Track.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. doi: 10.48550/ARXIV.2209.03003.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, March 2021. doi: 10.1038/s42256-021-00302-5.

François Mazé and Faez Ahmed. Diffusion models beat GANs on topology optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9108–9116, 2023. doi: 10.1609/aaai.v37i8.26093. URL https://ojs.aaai.org/index.php/AAAI/article/view/26093.

Zhenguo Nie, Tong Lin, Haoliang Jiang, and Levent Burak Kara. TopologyGAN: Topology optimization using generative adversarial networks based on physical fields over the initial domain. *Journal of Mechanical Design*, 143(3), February 2021. doi: 10.1115/1.4049533.

Sangeun Oh, Yongsu Jung, Seongsin Kim, Ikjin Lee, and Namwoo Kang. Deep generative design: Integration of topology optimization and generative models. *Journal of Mechanical Design*, 141 (11):111405, 2019. doi: 10.1115/1.4044229.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The Well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44989–45037. Curran Associates, Inc., 2024. doi: 10.52202/079017-1430. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4f9a5acd91ac76569f2fe291b1f4772b-Abstract-Datasets_and_Benchmarks_Track.html.

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77187–77200, 2023.

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*, pp. 234–241. Springer International Publishing, 2015. doi: 10.1007/978-3-319-24574-4_28.

Ole Sigmund and Kurt Maute. Topology optimization approaches. *Structural and Multidisciplinary Optimization*, 48(6):1031–1055, August 2013. doi: 10.1007/s00158-013-0978-6.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2210.07182.

Ronak Tali, Ali Rabeh, Cheng-Hau Yang, Mehdi Shadkhah, Samundra Karki, Abhisek Upadhyaya, Suriya Dhakshinamoorthy, Marjan Saadati, Soumik Sarkar, Adarsh Krishnamurthy, Chinmay Hegde, Aditya Balu, and Baskar Ganapathysubramanian. FlowBench: A large scale benchmark for flow simulation over complex geometries. *Data-centric Machine Learning Research*, 2(14): 1–35, 2025. URL https://data.mlr.press/volumes/02.html.

Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. Design-Bench: Benchmarks for data-driven offline model-based optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21658–21676, 2022. URL https://proceedings.mlr.press/v162/trabucco22a.html.

Gege Wen, Zongyi Li, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M. Benson. U-FNO–an enhanced Fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, May 2022. doi: 10.1016/j.advwatres.2022.104180.

Rebekka V. Woldseth, Niels Aage, J. Andreas Bærentzen, and Ole Sigmund. On the use of artificial neural networks in topology optimisation. *Structural and Multidisciplinary Optimization*, 65(10): 294, 2022. doi: 10.1007/s00158-022-03347-1.

Kentaro Yaji, Shintaro Yamasaki, and Kikuo Fujita. Data-driven multifidelity topology design using a deep generative model: Application to forced convection heat transfer problems. *Computer Methods in Applied Mechanics and Engineering*, 388:114284, 2022. doi: 10.1016/j.cma.2021.114284.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy T. Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary E. Ross, Katherine L. Bouman, and Yisong Yue. InverseBench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=U3PBITXNG6.

A Flow Physics and Numerical Benchmark Construction

This appendix specifies the flow problem, reference construction, parameter distributions, and numerical validation used throughout FluidTopoBench.

A.1 Governing Problem and Boundary Conditions

The normalized computational domain is $\Omega = [0, 1.5] \times [0, 1]$. Let $\rho : \Omega \to [0, 1]$ denote the fluid density field, where $\rho = 1$ indicates open fluid and $\rho = 0$ indicates solid. For a given operating condition c, the minimum-dissipation topology optimization problem is formulated as

$$\begin{aligned} \min_{\rho,\mathbf{u},p} \quad & \phi(\rho, \mathbf{u}) = \int_\Omega \left(\mu\, \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u}) + \frac{1}{2}\alpha(\rho)|\mathbf{u}|^2 \right) \mathrm{d}\mathbf{x}, \\ \text{s.t.} \quad & \rho_f(\mathbf{u}\cdot\nabla)\mathbf{u} - \nabla\cdot\big(2\mu\boldsymbol{\varepsilon}(\mathbf{u})\big) + \nabla p + \alpha(\rho)\mathbf{u} = 0 \quad \text{in } \Omega, \\ & \nabla\cdot\,\mathbf{u} = 0 \quad \text{in } \Omega, \\ & 0 \le \rho \le 1, \qquad \rho|_{\mathcal{O}} = 0, \\ & V(\rho) := \frac{1}{|\Omega|}\int_\Omega \rho \, \mathrm{d}\mathbf{x} \le V_f. \end{aligned} \tag{8}$$

Here $\mathcal{O} \subset \Omega$ denotes fixed solid obstacles and V_f represents the prescribed fluid-volume fraction budget. The design density ρ is discretized as piecewise constant (Q_0) on each quadrilateral cell, whereas the velocity field $\mathbf{u}$ and pressure field p are discretized using equal-order bilinear (Q_1–Q_1) finite elements.

Boundary conditions. Let Γ_{in}, Γ_{out}, and Γ_{w} denote the inlet, outlet, and impermeable wall segments of the boundary $\partial\Omega$, respectively. Inlet ports prescribe parabolic velocity profiles $\mathbf{u}_{\text{in}}$ normalized to discrete mean inflow speed $U_{\text{mean}} = 1$:

$$\mathbf{u} = \mathbf{u}_{\text{in}} \quad \text{on } \Gamma_{\text{in}}, \qquad \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_{\text{w}}. \tag{9}$$

At outlet boundaries, the static pressure and tangential velocity are set to zero,

$$p = 0, \qquad \mathbf{u}\cdot\mathbf{t} = 0 \quad \text{on } \Gamma_{\text{out}}, \tag{10}$$

where $\mathbf{t}$ is the outlet tangent. Normal velocity is left unconstrained in the weak momentum formulation. Fixed solid obstacles remain part of the computational mesh but are constrained to $\rho = 0$, suppressing fluid flow through Brinkman porosity penalization.

Flow and material parameters. We set the fluid mass density to $\rho_f = 1$. The characteristic length L_c is defined as the unweighted mean width of all grid-snapped physical inlets, yielding the dynamic viscosity

$$\mu = \frac{\rho_f U_{\text{mean}} L_c}{\mathrm{Re}}. \tag{11}$$

The Brinkman resistance is interpolated as

$$\alpha(\rho) = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min})\frac{1-\rho}{1+q_\alpha\rho}, \qquad \alpha_{\max} = \frac{2.5\mu}{0.01^2}, \qquad \alpha_{\min} = \frac{2.5\mu}{100^2}. \tag{12}$$

The limits $\alpha_{\max}$ and $\alpha_{\min}$ are the solid and fluid resistances; q_α controls the interpolation during optimization. Changing Re changes convection relative to the viscosity-scaled stress and drag terms while keeping the inlet-speed normalization fixed.

A.2 Reference Optimization and FE Reanalysis

Numerical solver. Reference designs are generated using the finite-element formulation of TopFlow (Alexandersen, 2023). The steady Navier–Stokes–Brinkman equations are discretized using equal-order bilinear Q_1–Q_1 elements with SUPG/PSPG streamline and pressure stabilization, and are solved using a damped Newton method. Design densities are updated via an optimality criteria (OC) method that enforces the volume fraction constraint V_f while maintaining $\rho|_{\mathcal{O}} = 0$. A multi-stage continuation scheme varies q_α during the design updates. No explicit density filters, sensitivity

filters, or perimeter penalties are applied. Accepted FE states satisfy the prescribed degree-of-freedom constraints, pass finite-value checks, and meet a relative residual tolerance of 10^{-6}. Design optimization stops at its objective-change criterion or iteration budget. A finite normal return under either stopping condition counts as completion.

Reference construction. Every physical instance initializes from a uniform feasible density field: $\rho_0(\mathbf{x}) = V_f$ on designable elements and $\rho_0(\mathbf{x}) = 0$ on $\mathcal{O}$. From this shared starting point, the pipeline branches into initial-response simulation and topology optimization:

$$\begin{aligned} &(c, V_f) \xrightarrow[\rho_0]{\text{uniform-design FE solve}} (\mathbf{u}_0, p_0), \\ &(c, V_f) \xrightarrow[\rho_0]{\text{topology optimization}} \rho_{\text{end}} \xrightarrow{\rho_{\text{end}} \geq 0.5} \rho_{\text{ref}} \xrightarrow{\text{binary FE reanalysis}} (\mathbf{u}_{\text{ref}}, p_{\text{ref}}, J_{\text{ref}}). \end{aligned} \tag{13}$$

The upper branch provides the initial response $(\mathbf{u}_0, p_0)$ for generation conditioning. The lower branch reaches the continuous endpoint ρ_{end} under the stated stopping policy, thresholds it at 0.5, and solves the FE state on ρ_{ref}. This last solve supplies both the reference fields and J_{ref}. The continuous endpoint is stored for the cold-start comparison.

Shared physical reanalysis. Reference and generated binary geometries use the same fixed-geometry FE analysis. Generation DPE, response errors, and candidate verification therefore use a shared solver protocol, with each response paired to its own input geometry.

Continuous initialization branch. Numerical initialization evaluates continuous generator outputs $\widehat{\rho}$ without binarization. These continuous fields initialize downstream optimization under the same optimizer, objective, and stopping criteria used for the cold-start baseline (initialized from ρ_0). Binary evaluation and continuous warm-start optimization use the same physical conditions but different design representations and endpoint comparators.

A.3 Sampling Distributions and Dataset Splits

FluidTopoBench systematically varies boundary layouts, flow regimes, volume budgets, obstacle placements, and spatial resolutions. Table 4 details the sampling parameter distributions for standard training/testing splits and targeted out-of-distribution (OOD) shifts; split sample sizes are listed in Table 1.

The five boundary families (dual passages, bends, T-junctions, Y-junctions, and multi-port layouts) have sampling weights $0.25, 0.25, 0.25, 0.10, 0.15$. Port perturbations follow template selection and respect separation and wall margins. Because HR-OOD-BC maintains identical pixel-level perturbation bounds on a 128×192 grid, its physical boundary variations are narrower than those of the base 64×96 resolution. Consequently, HR-OOD-BC represents a joint boundary-layout and resolution shift across independently sampled instances, rather than a paired grid-convergence ablation.

Each record in FluidTopoBench bundles operating parameters c, the uniform-design response $(\mathbf{u}_0, p_0)$, the thresholded binary geometry ρ_{ref}, its reanalyzed FE solution $(\mathbf{u}_{\text{ref}}, p_{\text{ref}}, J_{\text{ref}})$, and the continuous cold-start endpoint and associated optimization records. Validation checks split counts, metadata, port orientation, obstacle masks, and finite states. File schemas and parameter settings accompany the release.

Physical condition families. Figure 7 shows how port layouts and fixed solid regions define the available passages. The examples emphasize the Dual, Large, and Notch obstacle shifts alongside a held-out multi-port boundary layout. Each row pairs the actual operating layout and binary reference with six generators on that same instance; the rows are independently sampled conditions.

Table 4: **Physical parameter distributions and controlled out-of-distribution shifts.** Boundary-template coordinates and split manifests are provided in the data release.

Factor	Standard sampling	Controlled shift
Boundary layout	30 templates for training-range boundaries. The full catalog spans dual passages, bends, T/Y-junctions, and multi-port layouts.	10 held-out templates for OOD-BC and HR-OOD-BC; 40 templates in total.
Port perturbation	Port positions perturbed by up to ± 2 grid cells; widths perturbed by up to ± 1 cell, subject to minimum port separation and wall margins.	HR-OOD-BC applies the same pixel-level perturbation rule on a $2\times$ finer grid, yielding smaller physical displacements.
Reynolds number	Latin-hypercube sampling in $\log_{10}\mathrm{Re}$ over $\mathrm{Re} \in [10^{-3}, 200]$.	Low-Re uses $\mathrm{Re} \in [10^{-5}, 5 \times 10^{-4}]$.
Fluid-volume budget	$V_f \in \{0.25, 0.30, 0.35, 0.40, 0.45\}$.	Identical discrete levels preserved across all targeted shift splits.
Obstacle presence	60% single-obstacle configurations and 40% unobstructed flow domains.	Dual introduces two disjoint obstacles; Notch introduces a solid intrusion attached to the domain boundary.
Obstacle geometry	Circular, elliptical, and rectangular obstacles with equal probability; characteristic size parameter in $[0.12, 0.22]$.	Large scales the characteristic obstacle size parameter to $[0.24, 0.32]$.
Spatial resolution	64×96 quadrilateral cells.	HR-OOD-BC evaluates at 128×192 cells combined with held-out boundary templates.

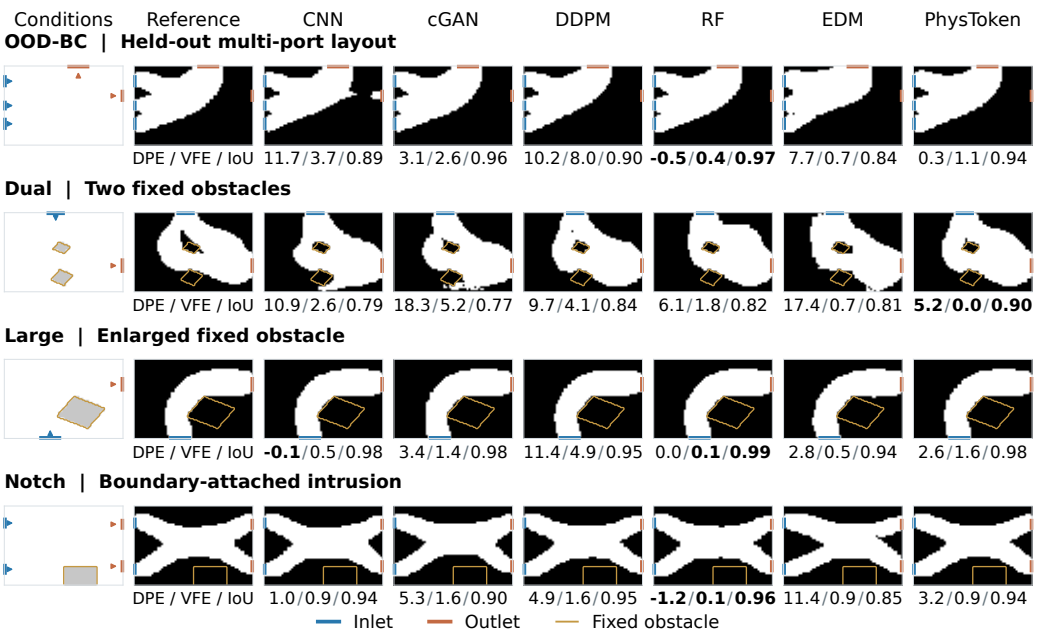


Figure 7: **Physical condition families and generated channels.** Blue and orange segments mark inlets and outlets; gold contours delimit fixed obstacles. White denotes fluid in the binary reference and generated designs. Under each generator, entries give DPE / VFE / IoU (the first two in percent), with the lowest DPE and VFE and highest IoU in bold within each row. Examples illustrate distinct routing conditions and model differences, rather than a matched parameter sweep or population-average performance.

A.4 Reference Quality and Numerical Validation

We evaluate reference quality through multi-start optimization at matched fluid volume, numerical reanalysis of fixed geometries, and state diagnostics across the full dataset. These checks assess both the competitiveness of the reference designs and the consistency of their physical targets.

Multi-start competitiveness. Each problem is optimized from five initializations: the uniform design, two smooth random perturbations, and two directional perturbations. All starts use the same optimizer and stopping rule. Each valid endpoint is binarized by retaining its highest-density designable cells to match the reference's fluid-cell count, then reanalyzed by FE. This holds the actual fluid volume fixed when comparing layouts.

Let $\mathcal{S}$ contain the starts with valid results and J_i^{mv} their matched-volume objectives. The best objective found and the additional reduction relative to the reference are

$$\begin{aligned} J_{\min} &= \min(\{J_{\mathrm{ref}}\} \cup \{J_i^{\mathrm{mv}} : i \in \mathcal{S}\}), \\ H &= 100\,\frac{J_{\mathrm{ref}} - J_{\min}}{J_{\mathrm{ref}}}. \end{aligned} \tag{14}$$

Including the reference gives $H \geq 0$; H measures the improvement found by this finite comparison. Since H is expressed in percent, $H \leq 1$ denotes at most 1% additional reduction. Failed starts are excluded from $\mathcal{S}$ and retained in coverage reporting.

The five-start comparison finds at most 1% additional dissipation reduction in 93.4% of assessed problems (Table 5). Small medians across all splits support the references as competitive baselines under the tested optimizer and initializations. The upper percentiles identify cases with more room for improvement.

Table 5: **Reference competitiveness at matched fluid volume.** H is the additional dissipation reduction found by five initializations. P95 denotes the 95th percentile; pooled statistics are computed over individual problems. Valid-start coverage is 20,725 of 20,750 starts.

		Reduction H (%)		Share (%)
Split	Problems	Median	P95	$H \leq 1$
Train	3,200	0.029	1.71	93.0
Validation	400	0.042	1.10	93.8
Test	200	0.018	1.05	94.5
OOD-BC	150	0.010	0.77	95.3
Low-Re	50	0.010	0.41	98.0
Dual	50	0.018	0.38	98.0
Large	50	0.003	0.60	98.0
Notch	50	0.031	0.93	94.0
Pooled	4,150	0.028	1.54	93.4

Fixed-geometry numerical stability. We preserve the binary geometry, physical parameters, and boundary trace while refining the 64×96 grid to 128×192; a nested subset also uses 256×384. Fine-grid solves use a relative residual tolerance of 10^{-9}. A separately assembled Taylor–Hood P_2/P_1 formulation provides an independent FE comparison under the same Brinkman model and boundary conditions.

Let J_{base} and J_{check} denote the baseline and comparison objectives. We measure absolute changes in reference dissipation and in the DPE of a generated/reference pair:

$$\delta_J = 100\,\frac{|J_{\mathrm{check}}(\rho_{\mathrm{ref}}) - J_{\mathrm{base}}(\rho_{\mathrm{ref}})|}{J_{\mathrm{base}}(\rho_{\mathrm{ref}})}, \tag{15}$$

$$\delta_{\mathrm{DPE}} = 100\,|\mathrm{DPE}_{\mathrm{check}} - \mathrm{DPE}_{\mathrm{base}}|. \tag{16}$$

Here δ_J is in percent and δ_{DPE} in percentage points. Each DPE uses the generated and reference objectives from the same analysis setting, preserving the paired design comparison.

Table 6: **Sensitivity of fixed-geometry evaluation.** Coverage lists references / generated–reference pairs. Grid refinement uses the 64×96 baseline; the independent formulation and solid-resistance checks use the 128×192 benchmark analysis as baseline. P95 denotes the 95th percentile.

		δ_J (%)		δ_{DPE} (pp)	
Check	Coverage	Median	P95	Median	P95
128×192 grid	4,150 / 440	1.017	3.072	0.026	0.800
256×384 grid	12 / 12	1.706	9.922	0.027	1.482
Independent P_2/P_1	24 / 24	0.681	13.885	0.016	0.895
$10\times$ solid resistance	4,150 / 440	7.267	30.222	0.129	3.386

Doubling each grid dimension changes reference dissipation by at most 5% in 97.5% of assessed cases. Small paired-DPE changes also support the stability of relative design evaluation (Table 6). The independent formulation shows close median agreement, with larger reference-objective differences in the upper tail.

Increasing solid resistance tenfold produces a larger change in reference dissipation than grid refinement. This check measures sensitivity to the porous-solid material model; reference competitiveness and the other numerical comparisons use the specified benchmark resistance.

Full-dataset state consistency. All 40,500 stored binary-reference states satisfy the 10^{-6} algebraic residual tolerance. Direct quadrature of the saved nodal states reproduces every stored objective to within 6.7×10^{-15} relative difference. Boundary-flux imbalance has a median of 1.38% and a 95th percentile of 2.20%. For the dimensionless strong-divergence diagnostic $D = |\Omega|^{1/2}\|\nabla\mathbf{u}\|_{L_2}/\|\mathbf{u}\|_{L_2}$, the corresponding values are 0.916 and 1.318. These conservation diagnostics complement the algebraic residual at the retained stabilized discretization.

B Baseline Architectures and Training Protocols

We report model architectures, conditioning, and training budgets below. Layerwise settings and normalization constants are provided with the model configurations.

B.1 Topology Generators

Input and conditioning. All generators predict a topology on the 64×96 training grid and share the same physical-condition interface. Initial-response conditioning uses the uniform-design pressure and velocity fields together with V_f, Re, and obstacle-distance information. The boundary-map variant replaces the response fields with explicit inlet/outlet distance maps while retaining the remaining conditions. Both representations encode the boundaries, while differing in response information and boundary encoding. The initial-response input requires a uniform-design FE solve, reused across candidates under the same conditions. Response fields use training-set normalization; pressure is gauge-centered and scaled by Re first.

Architectures. All six generators use a common residual U-Net family with base width 128, channel multipliers $(1, 2, 4, 8)$, and two residual blocks per level. The backbone uses GroupNorm, SiLU activations, and attention at coarse spatial scales. The baselines differ mainly in their generative formulation: CNN performs deterministic topology prediction with BCE and Dice losses; cGAN adds a PatchGAN discriminator with hinge adversarial training; DDPM predicts diffusion noise; RF predicts the velocity of a noise-to-topology flow; and EDM uses preconditioned denoising. PhysToken extends the diffusion backbone with 16 physical-condition tokens and scalar modulation by Re and V_f. Complete layerwise settings are provided in the released configurations.

Training and inference. All generators are optimized with AdamW using BF16 computation and gradient clipping. The generator learning rate is 10^{-4}; the cGAN discriminator uses 5×10^{-5}. Diffusion and flow models use weight decay 10^{-2}, learning-rate warmup, and EMA checkpoints. Checkpoints are selected using the corresponding validation objective—reconstruction loss for CNN/cGAN and the noise, flow, or denoising objective for the stochastic models. FE performance on

the test set is never used for checkpoint selection. Model-specific budgets and inference procedures are summarized in Table 7.

Table 7: **Generator training and inference budgets.** Parameter counts refer to the generator; the cGAN count excludes its discriminator.

Generator	Epochs	Updates	Batch	Parameters	Inference
CNN	80	20,000	128	138.5M	One forward
cGAN	60	15,000	128	138.5M	One forward
DDPM	200	50,000	128	142.9M	DDIM, 50 steps
RF	200	50,000	128	142.9M	Euler, 50 steps
EDM	200	50,000	128	142.9M	EDM Heun, 50 steps
PhysToken	200	50,000	128	156.9M	DDIM, 50 steps

B.2 Response Predictors

Input and targets. Each response predictor receives the binary geometry together with the volume budget, Reynolds number, obstacle-distance information, and boundary labels, and predicts pressure and vector velocity on the same geometry. Pressure training targets are gauge-centered, multiplied by Re, and divided by a training-set scale; the two velocity components share a training-set scale. All reported physical metrics are computed after returning predictions to physical units.

Architectures. The response suite spans convolutional, spectral, and branch–trunk operator families. Table 8 records the architecture-level choices needed to distinguish the six predictors; lower-level layer and operator settings are provided in the released model configurations.

Table 8: **Response-predictor architectures.** PINO shares the U-FNO backbone and differs through its physics-informed training objective.

Predictor	Architecture summary
U-Net	U-Net with a ResNet-18 encoder.
FNO	Four Fourier spectral layers with width 64.
U-FNO	Six spectral blocks augmented with local U-Net paths.
PINO	Same U-FNO backbone.
CNO	Three convolutional-operator levels with residual blocks.
DeepONet	Convolutional geometry branch with a Fourier-feature coordinate trunk.

Training and checkpoint selection. All predictors are trained against the matched FE fields on binary geometries. The common validation criterion is channelwise relative-L_2 data error, which selects the checkpoint for every predictor, including PINO. PINO augments its data-fitting objective with divergence, momentum, solid-region, and smoothness penalties:

$$\mathcal{L} = \mathcal{L}_{\mathrm{data}} + 10^{-2}\left(10^{-1}\mathcal{L}_{\mathrm{div}} + 10^{-3}\mathcal{L}_{\mathrm{mom}} + 10^{-1}\mathcal{L}_{\mathrm{solid}} + 10^{-4}\mathcal{L}_{\mathrm{smooth}}\right). \tag{17}$$

The momentum term contains convection, symmetric-gradient viscosity, pressure, and Brinkman drag with coefficients determined by each physical instance. Table 9 gives the model-specific optimization schedules.

Table 9: **Predictor training settings.** Every checkpoint is selected by the same validation channelwise relative-L_2 data loss.

Predictor	Epochs	Batch	Learning rate	Weight decay	Precision
U-Net	400	256	10^{-4}	0	BF16
FNO	400	256	5×10^{-4}	0	FP32
U-FNO	300	128	5×10^{-4}	0	FP32
PINO	300	128	5×10^{-4}	0	FP32
CNO	400	128	5×10^{-4}	0	FP32
DeepONet	400	512	5×10^{-4}	10^{-4}	BF16

C Evaluation Metrics and Statistical Protocols

Table 10 distinguishes binary FE targets from continuous optimization endpoints. Metric formulas use dimensionless ratios; reported percentages multiply these ratios by 100.

Table 10: Symbols and evaluation targets. Percentages are displayed ratios multiplied by 100.

Symbol	Meaning
c	Boundary conditions, flow parameters, budget and fixed obstacles.
$\widehat{\rho}, \rho_{\mathrm{end}}$	Floating-point generator output and continuous reference-optimization endpoint.
$\rho, \rho_{\mathrm{ref}}$	Binary evaluation and reference geometries; ρ also denotes a continuous variable when explicitly used in the optimization problem.
$J_c(\rho), J_{\mathrm{ref}}$	Native FE objective of the specified geometry and binary reference.
$\phi_{P,\mathrm{end}}, \phi_{\mathrm{cold}}$	Continuous warm and cold optimization endpoint objectives.
$a_{\rho,h}, \Pi_h$	Common discrete dissipation form and fixed cell-to-node interpolation.
η_ρ, E_J	Dissipation-weighted field error and error relative to native FE dissipation.
γ, ϵ	Required binary-reference improvement and allowed continuous cold-endpoint excess; both are dimensionless.

C.1 Response Errors and Dissipation

Field error measures reconstruction, η_ρ measures velocity error in the dissipation norm, and E_J measures the scalar objective mismatch. They can therefore assign different rankings to the same predictors.

State reconstruction. Velocity error uses the full vector field. Pressure error uses the target gauge specified by the data representation; predictions are not recentered after inference to reduce error. The outlet condition fixes the FE pressure reference. Gradient and divergence diagnostics provide additional field checks.

Dissipation-weighted error. The fixed cell-to-node operator Π_h maps both predicted and target velocities to the Q1 representation used for quadrature. We define the symmetric bilinear form

$$a_{\rho,h}(\mathbf{v}, \mathbf{w}) = \int_\Omega \left[2\mu\boldsymbol{\varepsilon}(\Pi_h\mathbf{v}) : \boldsymbol{\varepsilon}(\Pi_h\mathbf{w}) + \alpha(\rho)(\Pi_h\mathbf{v}) \cdot (\Pi_h\mathbf{w})\right] \mathrm{d}\mathbf{x}. \tag{18}$$

For the velocity residual $\mathbf{r} = \widehat{\mathbf{u}} - \mathbf{u}$, the relative energy-norm error is defined by

$$\eta_\rho^2 = \frac{a_{\rho,h}(\mathbf{r}, \mathbf{r})}{a_{\rho,h}(\mathbf{u}, \mathbf{u})}. \tag{19}$$

Unlike standard L_2 errors, η_ρ weights velocity residuals by the local viscous dissipation and Brinkman drag.

Relation to objective error. Let $\Phi_{\rho,h}(\mathbf{v}) = \frac{1}{2}a_{\rho,h}(\mathbf{v}, \mathbf{v})$ denote the discrete dissipation functional. Expanding the quadratic form about the ground-truth velocity $\mathbf{u}$ yields

$$\frac{\Phi_{\rho,h}(\widehat{\mathbf{u}}) - \Phi_{\rho,h}(\mathbf{u})}{\Phi_{\rho,h}(\mathbf{u})} = \frac{2a_{\rho,h}(\mathbf{u}, \mathbf{r})}{a_{\rho,h}(\mathbf{u}, \mathbf{u})} + \eta_\rho^2, \tag{20}$$

$$\left|\frac{\Phi_{\rho,h}(\widehat{\mathbf{u}}) - \Phi_{\rho,h}(\mathbf{u})}{\Phi_{\rho,h}(\mathbf{u})}\right| \le 2\eta_\rho + \eta_\rho^2. \tag{21}$$

The cross term depends on alignment in the dissipation bilinear form. Equal η_ρ can therefore accompany different objective errors. The identity and bound compare the shared interpolated quadratic $\Phi_{\rho,h}$. In contrast, E_J compares $\widehat{J} = \Phi_{\rho,h}(\widehat{\mathbf{u}})$ with $J_c(\rho)$ computed from the native FE nodal state. Reconstructing that state from cell-centered targets need not preserve its quadratic exactly. Both targets refer to the same binary geometry.

Spatial error decomposition. The energy-norm error can be localized by restricting the bilinear form in Eq. (18) to each element Ω_e:

$$w_e = \frac{a_{\rho,h}^{(e)}(\mathbf{r}, \mathbf{r})}{a_{\rho,h}(\mathbf{u}, \mathbf{u})}, \qquad \sum_e w_e = \eta_\rho^2. \tag{22}$$

Figure 4(a) visualizes the spatial distribution of w_e alongside standard velocity residuals. The map sums to η_ρ^2. Signed objective changes also include the cross term in Eq. (20), so w_e is not an additive map of E_J.

C.2 Aggregation and Uncertainty

Physical instances are the bootstrap units. Predictions, candidate samples, and optimization attempts from the same instance remain together under resampling. Table 11 separates this resampling unit from the weights used to report each result.

Table 11: **Averaging and failure handling.** Bootstrap resampling uses physical instances; reporting weights depend on the evaluation.

Analysis	Reported summary	Coverage / pairing
Generation and reference response	Instances have equal weight. Generator samples are averaged within each instance first.	Field and objective errors require valid targets. Deterministic copies add no independent observations.
Candidate selection	Passing is computed per generator–split condition; macro results weight these conditions equally.	Passing retains all planned pools. Regret conditions on FE-valid, volume-eligible selections.
Numerical initialization	Joint rates retain the original attempt weights.	All scheduled pairs, including failures, remain in joint rates. Bootstrap sampling is stratified by split.

Bootstrap intervals. Reported intervals are pointwise 95% physical-instance bootstrap intervals. They describe uncertainty across sampled instances at the evaluated checkpoints and generation realizations, not variability from retraining. Generation summaries average samples within each instance before resampling; medians are then recomputed across the resampled instances. Generated-response macro results weight generator–split paired means equally (Appendix F.2).

Paired errors and joint outcomes. Paired field and objective errors use instances where both required values are valid. On the same instances with the same weights, the mean paired difference equals the difference of means; separately reported marginal means may use different valid populations. Medians do not generally commute with differencing. Joint passing and completion comparisons instead retain all planned pairs, including failures. Valid error values are never assigned to failed FE solves; coverage is recorded separately.

Qualitative examples. Shared-instance examples illustrate model differences, metric disagreement, or verified improvement. Quantitative claims, including outcome frequencies, use the evaluation population specified for each analysis. Selection rules and source mappings accompany the figure assets.

D Generation and Conditioning Results

Table 12 extends the main generation comparison to all six standard evaluation splits. Stochastic samples are averaged within each physical instance before split-level aggregation, following Appendix C.2. Entries are point estimates under default initial-response conditioning; the result records support the uncertainty calculations in Appendix C.2.

Table 12: **Topology-generation performance across all standard splits.** DPE and VFE are percentages and IoU is a ratio. Median DPE is computed across physical instances after averaging stochastic samples within each instance. Bold marks the most favorable point estimate within each split and metric.

Split	Metric	CNN	cGAN	DDPM	RF	EDM	PhysToken
Test	Mean DPE (%) ↓	0.72	2.27	8.21	**0.18**	2.75	2.22
	Median DPE (%) ↓	0.09	0.34	4.33	**-0.22**	0.09	0.63
	VFE (%) ↓	1.35	1.81	4.14	**1.07**	2.39	1.53
	IoU ↑	**0.961**	0.945	0.899	0.946	0.864	0.929
OOD-BC	Mean DPE (%) ↓	2.23	3.26	9.33	**1.77**	7.05	3.51
	Median DPE (%) ↓	0.43	0.79	3.10	**0.02**	1.02	1.00
	VFE (%) ↓	1.55	2.47	4.02	**1.13**	2.65	1.58
	IoU ↑	**0.922**	0.910	0.887	0.921	0.856	0.906
Low-Re	Mean DPE (%) ↓	**-1.38**	1.61	6.11	-0.74	-1.02	5.98
	Median DPE (%) ↓	-0.74	0.09	3.65	-0.41	**-1.22**	2.18
	VFE (%) ↓	1.00	1.61	3.63	**0.68**	2.25	1.58
	IoU ↑	**0.969**	0.957	0.913	0.956	0.890	0.892
Dual	Mean DPE (%) ↓	**5.20**	10.35	8.54	5.23	11.79	5.86
	Median DPE (%) ↓	0.66	0.69	4.54	**0.07**	2.03	1.11
	VFE (%) ↓	1.72	2.15	4.22	**1.33**	2.59	1.72
	IoU ↑	0.907	0.898	0.869	**0.914**	0.840	0.893
Large	Mean DPE (%) ↓	2.06	5.56	8.39	**1.52**	8.56	3.96
	Median DPE (%) ↓	0.42	0.99	3.86	**0.10**	0.88	0.66
	VFE (%) ↓	1.68	2.37	4.46	**1.47**	2.72	1.76
	IoU ↑	**0.940**	0.924	0.891	0.938	0.860	0.916
Notch	Mean DPE (%) ↓	2.48	1.88	8.31	**-1.14**	-0.18	1.19
	Median DPE (%) ↓	0.26	0.60	2.83	-0.30	**-0.58**	0.58
	VFE (%) ↓	1.40	1.87	3.68	**0.97**	2.41	1.42
	IoU ↑	0.945	0.936	0.909	**0.950**	0.881	0.930

The full split comparison preserves the distinction between geometric and physical quality. RF has the lowest mean VFE on every standard split and the lowest mean DPE on four of six splits, whereas CNN has the highest IoU on Test, OOD-BC, Low-Re, and Large. Mean and median DPE distinguish average objective error from its central value across instances.

D.1 Initial-Response Conditioning

Figure 3 reports paired physical-instance uncertainty for Test and OOD-BC. Table 13 extends the point estimates to all six standard splits. For each physical instance, stochastic samples are averaged under boundary-map (BM) and initial-response (IR) conditioning. The paired difference $\Delta_{\mathrm{DPE}} = \mathrm{DPE}_{\mathrm{BM}} - \mathrm{DPE}_{\mathrm{IR}}$ is averaged over jointly valid instances. Positive values favor initial-response conditioning.

Initial-response conditioning has its clearest benefit on OOD-BC, the only split whose point estimates favor it for all six generators. On other splits, the direction depends on both generator and physical setting.

Table 13: **Paired mean-DPE effect of initial-response conditioning across all standard splits.** Entries are paired point estimates in percentage points. Positive values favor initial-response conditioning.

Split	CNN	cGAN	DDPM	RF	EDM	PhysToken
Test	-0.69	-2.14	-6.30	+0.26	+21.45	-0.25
OOD-BC	+12.65	+20.89	+28.12	+9.41	+80.53	+8.72
Low-Re	+1.10	-3.46	-2.66	+0.14	+11.79	-4.42
Dual	+1.79	-0.68	-0.92	-1.23	+33.16	+1.73
Large	-0.32	-3.58	-5.88	+0.35	+25.68	-0.12
Notch	-2.47	-1.70	-8.05	+1.74	+24.38	-0.59

E Controlled Geometric Perturbations

This analysis tests whether equal geometric scores imply equal dissipation. Paired edits hold the changed-cell count, fluid volume, and reference IoU fixed, while an additional matching control restricts local geometric differences. The resulting FE comparisons examine information not captured by the geometric metrics used in the generation benchmark.

E.1 Equal-Budget Edit Construction

Equal geometric budget. Starting from a binary reference geometry with N_f fluid cells, each intervention removes k fluid cells and adds k cells in a common compensation region, with $k \in \{4, 12\}$. Ports, fixed obstacles, and exterior boundaries are protected. Every high/low pair therefore has identical fluid volume, exactly $2k$ changed cells, and the same overlap with the reference,

$$\mathrm{IoU} = \frac{N_f - k}{N_f + k}. \tag{23}$$

The compensation region is shared within an instance and budget, so any difference between the two edited designs cannot be attributed to edit size, volume mismatch, or reference IoU.

Edit families. Removal regions are chosen from the unmodified reference geometry using its strain-based activity proxy. High- and low-activity candidates are drawn from the upper and lower activity quartiles, respectively, before either edited geometry is reanalyzed. Five fixed operators vary only the spatial arrangement of the removed cells: *compact* edits form a local patch, *streamwise* and *cross-stream* edits elongate the patch along and across the local flow direction, *two-site* edits split the budget between two separated patches, and *distributed* edits use separated boundary cells. Exact operator definitions and construction coordinates are included in the released perturbation records.

Geometry matching. Equal edit budgets do not by themselves equalize the surrounding channel geometry. We therefore form a matched comparison that additionally constrains local channel width and distances to ports and obstacles, then pairs high- and low-activity candidates with nearby geometric features. Matching controls these local descriptors while allowing other geometric differences between the two neighborhoods.

E.2 Location Effects and Matching Control

For a matched high/low pair, the reported contrast is

$$\Delta J_{\mathrm{HL}} = 100\,\frac{J_c(\rho_{\mathrm{high}}) - J_c(\rho_{\mathrm{low}})}{J_c(\rho_{\mathrm{ref}})} \quad \text{percentage points.} \tag{24}$$

Positive values mean that the high-activity edit has higher dissipation. Figure 8(a) shows that the effect is strongly family-dependent at the larger edit budget: compact, streamwise, cross-stream, and distributed edits have positive mean contrasts, whereas the two-site construction reverses sign. Thus equal-volume, equal-overlap errors need not have equal physical consequences.

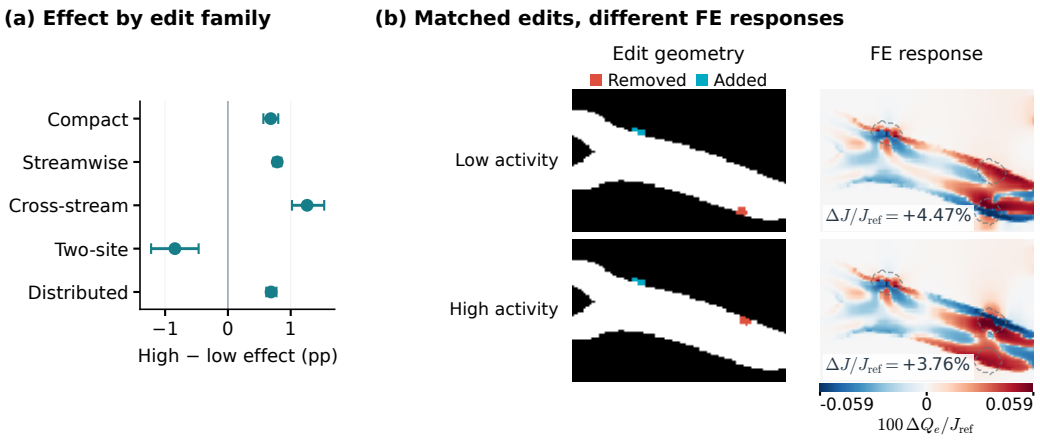


Figure 8: **Equal geometric scores do not determine dissipation.** (a) Mean high-minus-low contrasts for matched, route-preserving edits ($k = 12$), with pointwise 95% instance-bootstrap intervals. (b) Two equal-volume, equal-overlap edits and their FE dissipation changes. Dashed contours mark the common four-cell neighborhood. The example illustrates a reversed activity ordering; panel (a) reports the population comparison.

Table 14 compares matched and unmatched edits on their common available instances, requiring all four edited geometries to retain the flow route. This intersection can be smaller than the matched population in Figure 8(a). The table reports point estimates; counts and paired intervals accompany the result records.

Table 14: **Effect of local geometric matching on the high-minus-low objective contrast.** Values are percentage points on the common subset for which both matched and unmatched pairs are available and retain the required flow route.

	$k = 4$		$k = 12$	
Construction	Unmatched	Matched	Unmatched	Matched
Compact	0.711	0.501	2.094	0.682
Streamwise	0.534	0.442	1.287	0.785
Cross-stream	1.163	0.113	5.253	1.258
Two-site	1.156	-0.685	3.545	-0.844
Distributed	0.348	0.315	0.876	0.687

Matching lowers the signed mean contrast for every construction. For two-site edits, it reverses the sign at both budgets. Physical consequences remain sensitive to the location and arrangement of errors with equal pixel, volume, and overlap budgets.

E.3 Non-Local Dissipation Response

The FE objective admits an elementwise decomposition,

$$Q_e = \int_{\Omega_e} \left[\mu \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u}) + \tfrac{1}{2}\alpha(\rho)|\mathbf{u}|^2\right] \mathrm{d}\mathbf{x}, \qquad \sum_e Q_e = J_c(\rho). \tag{25}$$

For a radius-r neighborhood $\mathcal{N}_r$ of the changed cells, including removal and compensation, the outside share is

$$S_{\mathrm{out}}(r) = \frac{\sum_{e \notin \mathcal{N}_r} \left|Q_e^{\mathrm{edit}} - Q_e^{\mathrm{ref}}\right|}{\sum_e \left|Q_e^{\mathrm{edit}} - Q_e^{\mathrm{ref}}\right|}. \tag{26}$$

Across construction families, edit budgets, and matching settings, the mean outside share is 38.23–62.99% for a four-cell neighborhood and 25.48–43.88% for an eight-cell neighborhood. Local topology edits therefore produce substantial dissipation redistribution away from the edited region.

This is a share of absolute change, with no area normalization; it is distinct from the signed contribution to the total objective change. The paired maps in Figure 8(b) use a common neighborhood of both removal regions and their shared compensation cells.

In Figure 8(b), both edits increase dissipation, by 4.47% and 3.76% relative to the same reference. Their high-minus-low contrast is therefore -0.71 percentage points despite identical volume and IoU. This example illustrates a reversed activity ordering, not the magnitude or frequency of effects across the dataset. The paired population results are reported in Figure 8(a).

F Response Prediction and Generalization

F.1 Reference Geometries and Metric Comparisons

Table 15 complements the Test and OOD-BC results in Table 3 with the remaining standard splits. Predictor rankings vary with both physical setting and error metric. On Notch, PINO has the lowest pressure, velocity, and dissipation-weighted errors, whereas U-FNO has the lowest objective error. On Dual and Large, CNO leads in velocity and objective accuracy, while U-FNO has the lowest weighted error.

Table 15: **Reference-state prediction under additional physical shifts.** Entries are mean dimensionless errors over physical instances; lower is better. Bold marks the lowest point estimate in each row.

Split	Metric	U-Net	FNO	U-FNO	PINO	CNO	DeepONet
Low-Re	$E_p \downarrow$	0.0620	0.0433	**0.0241**	0.0254	0.0392	0.0575
	$E_u \downarrow$	0.0290	0.0220	0.0102	**0.0099**	0.0154	0.0679
	$\eta_\rho \downarrow$	0.1580	0.0871	**0.0332**	0.0335	0.0488	0.6007
	$E_J \downarrow$	0.0515	0.0416	0.0384	0.0393	**0.0374**	0.4615
Dual	$E_p \downarrow$	0.1520	0.1210	0.0940	0.0935	**0.0738**	0.1346
	$E_u \downarrow$	0.0656	0.0539	0.0339	0.0343	**0.0285**	0.1511
	$\eta_\rho \downarrow$	0.2148	0.1439	**0.0723**	0.0744	0.0729	1.0624
	$E_J \downarrow$	0.1171	0.0644	0.0580	0.0574	**0.0484**	1.5626
Large	$E_p \downarrow$	0.1302	0.0996	0.0723	0.0714	**0.0678**	0.1137
	$E_u \downarrow$	0.0519	0.0463	0.0270	0.0265	**0.0237**	0.1194
	$\eta_\rho \downarrow$	0.1903	0.1358	**0.0635**	0.0639	0.0678	0.9038
	$E_J \downarrow$	0.0651	0.0538	0.0493	0.0483	**0.0435**	1.1076
Notch	$E_p \downarrow$	0.0941	0.3067	0.0721	**0.0618**	0.1050	0.0830
	$E_u \downarrow$	0.0371	0.0659	0.0197	**0.0190**	0.0203	0.0818
	$\eta_\rho \downarrow$	0.1893	0.1993	0.0563	**0.0540**	0.1508	0.6706
	$E_J \downarrow$	0.1211	0.1581	**0.0482**	0.0493	0.0865	0.5857

Within-instance ranking. For metrics $m, n \in \{E_u, \eta_\rho, E_J\}$, let $m_{i,a}$ denote the error of predictor a on physical instance i. We measure the fraction of predictor pairs whose ordering reverses between metrics:

$$r_i(m,n) = \frac{1}{15} \sum_{a<b} \mathbf{1}[(m_{i,a} - m_{i,b})(n_{i,a} - n_{i,b}) < 0]\,. \tag{27}$$

Each instance contributes all 15 predictor pairs; exact ties count as non-reversals. Figure 4(b) reports equally weighted instance means and pointwise 95% intervals from 2,000 instance-bootstrap resamples, keeping all predictor comparisons from one instance together. On Test, E_u–η_ρ disagreement is 6.97%, compared with 33.00% for E_u–E_J and 33.97% for η_ρ–E_J. Comparisons involving E_J show more reversals in every standard split. The statistic records ordering changes without weighting their error magnitudes.

Spatial examples under obstacle shifts. Figure 9 compares all six predictors on shared binary references from Dual, Large, and Notch. These cases complement the Test and OOD-BC examples in the main text and expose local differences that are not apparent from scalar errors alone.

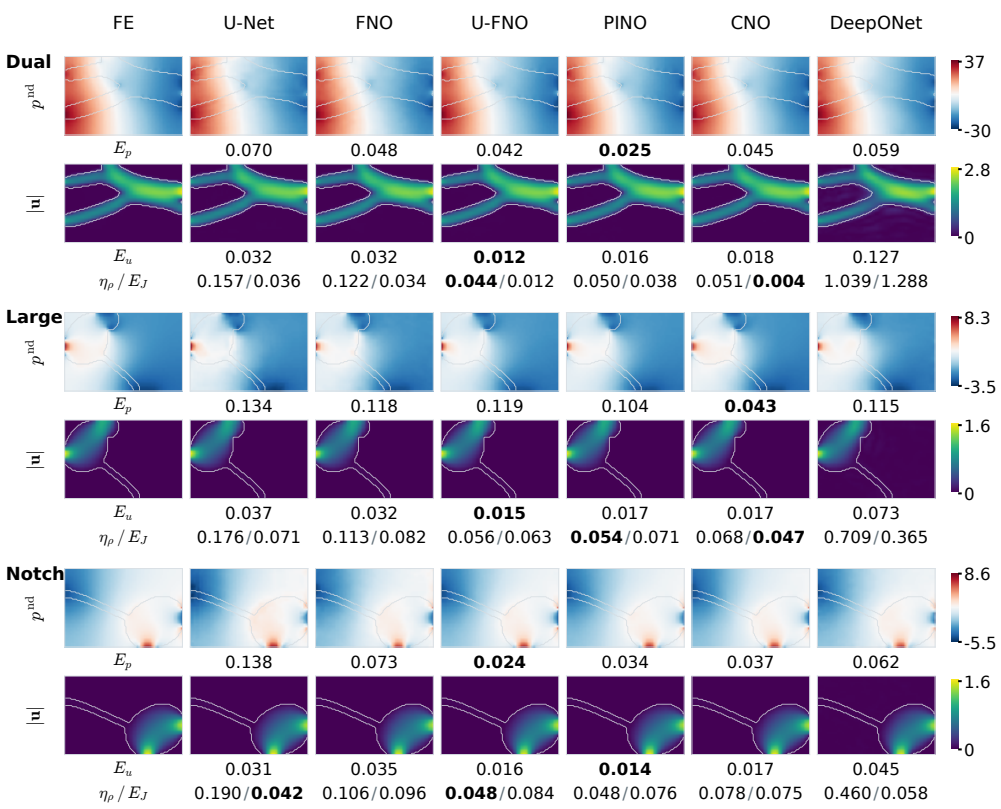


Figure 9: **Six-predictor responses under obstacle shifts.** Each split uses one independently sampled reference geometry. Rows show gauge-centered pressure multiplied by Re and physical velocity magnitude, with a common color scale across the FE target and all predictors. Labels report E_p and $E_u/\eta_\rho/E_J$; the lowest unrounded value is bold. Examples are selected near the split-wise median velocity error with geometric variety.

F.2 Generated Geometries

The same frozen predictor is evaluated on reference and generated geometries under matched operating conditions, with an FE target for each input geometry. Valid paired sampling differences are averaged within each physical instance, then aggregated by generator and split. Table 16(a) weights the resulting generator–split means equally. Missing FE targets are reported in coverage and excluded from paired error means.

Mean velocity error increases for all six predictors in this macro comparison. Weighted and objective errors increase for five predictors but decrease for DeepONet, whose reference-state errors are comparatively large. A negative generated-minus-reference difference therefore indicates improvement relative to that predictor's own reference-state error, not lower absolute error than other predictors.

F.3 Resolution Transfer

Native inference operates on the HR grid. The wrapper downsamples inputs to the training grid, applies the same checkpoint, and upsamples predictions; both paths are evaluated on the same HR instances against their HR FE targets. HR-OOD-BC combines boundary-layout and resolution changes (Appendix A.3).

Table 16(b) shows lower velocity error for every predictor under the wrapper. CNO improves on all three metrics, while DeepONet's lower velocity error accompanies higher weighted and objective errors. Inference scale therefore affects field reconstruction and dissipation prediction differently. Per-

condition results, coverage, and full bootstrap intervals are provided in the accompanying evaluation assets.

Table 16: **Prediction under generated-geometry and resolution shifts. (a)** Generated-minus-reference error, equally averaged over generator–split paired means; positive values indicate higher generated-input error. **(b)** Mean errors on the same HR-OOD-BC instances with fixed checkpoints; bold marks the lower value within each model–metric pair. All values are dimensionless.

Metric / inference path		U-Net	FNO	U-FNO	PINO	CNO	DeepONet
(a) Generated geometries: paired error change							
ΔE_u		$+0.0106$	$+0.0084$	$+0.0103$	$+0.0101$	$+0.0102$	$+0.0072$
$\Delta \eta_\rho$		$+0.0125$	$+0.0141$	$+0.0156$	$+0.0152$	$+0.0140$	-0.0114
ΔE_J		$+0.0041$	$+0.0104$	$+0.0095$	$+0.0088$	$+0.0095$	-0.2800
(b) HR-OOD-BC: native and wrapper inference							
$E_u \downarrow$	Native	0.6994	0.6341	0.5616	0.5763	0.5737	0.8915
	Wrapper	**0.4088**	**0.3373**	**0.1390**	**0.1426**	**0.1041**	**0.6577**
$\eta_\rho \downarrow$	Native	1.1261	0.6998	0.8125	0.7753	1.8050	**4.1723**
	Wrapper	**0.6429**	**0.4853**	**0.3045**	**0.3109**	**0.2603**	4.5047
$E_J \downarrow$	Native	0.6514	0.7135	0.3894	0.3809	6.7810	**17.2794**
	Wrapper	**0.3089**	**0.2614**	**0.1678**	**0.1740**	**0.1097**	27.4306

G Candidate Selection and Reference Improvement

G.1 Fixed-Pool Selection

Eligibility and verification. For each physical instance, all selectors receive the same five-candidate pool $\mathcal{C}(c)$ from a stochastic generator. The unfiltered Uniform baseline samples the full pool; all other rules use the observable volume-eligible set

$$\mathcal{E}(c) = \{\rho \in \mathcal{C}(c) : \mathrm{VFE}(\rho) \le \tau\}, \qquad \tau = 0.03.$$

A candidate passes the default quality band only after FE verification:

$$\mathcal{P}(c) = \{\rho \in \mathcal{E}(c) : \text{FE valid},\ \mathrm{DPE}(\rho) \le \delta\}, \qquad \delta = 0.10. \tag{28}$$

Every planned pool remains in the passing-rate denominator; absent selections and failed FE verification are therefore non-passing. Predictors rank candidates in $\mathcal{E}(c)$ using $\widehat{J}$ from geometry and operating conditions, without candidate FE labels. FE screening instead evaluates every candidate in $\mathcal{E}(c)$.

Conditional regret. For a pool containing at least one FE-valid eligible candidate, define

$$J_{\mathrm{best}}(c) = \min_{\substack{\rho \in \mathcal{E}(c) \\ \text{FE valid}}} J_c(\rho).$$

For a volume-eligible selected candidate ρ_π with valid FE analysis,

$$\mathrm{Regret}_\pi(c) = 100\, \frac{J_c(\rho_\pi) - J_{\mathrm{best}}(c)}{J_{\mathrm{best}}(c)}. \tag{29}$$

Failed, absent, or ineligible selections are excluded from regret and remain non-passing. The reported P95 averages conditionwise percentiles. For selectors operating on the same pools and weights,

$$1 - Q_\pi = \underbrace{1 - Q_{\mathrm{FE}}}_{\text{no passing candidate}} + \underbrace{Q_{\mathrm{FE}} - Q_\pi}_{\text{passing candidate missed}}. \tag{30}$$

Here $Q_\pi \in [0,1]$ is the passing probability under selector π and Q_{FE} is its same-pool FE-screening counterpart.

Volume filtering improves on unfiltered uniform selection, and all six predictors add a passing-rate gain over maximum volume (Table 17). Among predictors, CNO has the highest passing rate and lowest P95 regret; DeepONet has the lowest mean regret.

CNO's macro gain remains positive across $\tau \in \{0.01, 0.03, 0.05\}$ and $\delta \in \{0.05, 0.10, 0.20\}$ (0.517–3.909 pp). The full grid and condition-level intervals are included in the evaluation records.

Table 17: **Selection quality in shared five-candidate pools.** Macro results weight generator–split conditions equally. Passing includes all pools; regret uses valid eligible selections, and P95 averages conditionwise percentiles. Gain is relative to maximum volume; bold marks the best point estimate excluding FE screening.

	Pool passing		Conditional regret	
Rule	Rate (%) ↑	Gain (pp) ↑	Mean (%) ↓	P95 (%) ↓
Uniform	71.329	-18.117	3.312	11.686
Eligible uniform	89.109	-0.337	3.193	11.707
Minimum VFE	89.293	-0.153	3.114	11.386
Maximum volume	89.446	0.000	1.928	8.489
U-Net	91.442	1.997	1.293	2.790
FNO	91.331	1.885	1.402	3.293
U-FNO	91.658	2.213	1.040	1.796
PINO	91.678	2.232	1.051	1.775
CNO	**91.947**	**2.501**	0.833	**1.162**
DeepONet	91.830	2.384	**0.828**	2.727
FE screening	92.561	3.115	0.000	0.000

G.2 Reference Improvement and RF Proposals

Reference improvement requires valid FE analysis and the original constraints in Eq. (7):

$$V(\rho) \le V_f, \qquad \rho|_{\mathcal{O}} = 0, \qquad J_c(\rho) \le (1-\gamma) J_{\mathrm{ref}}. \tag{31}$$

At $\gamma = 0$ the selected design matches or improves upon the binary reference; positive γ requires an improvement margin. Eligibility uses the volume cap here, replacing the default VFE band.

Within five-candidate pools, CNO and maximum volume are close at reference parity by point estimate; CNO selects more designs meeting positive margins (Table 18(a)). RF16+CNO uses sixteen RF proposals per instance, clears fixed-obstacle cells, and ranks the volume-eligible binary candidates. The chosen design is FE-verified; all planned instances remain in the split denominator.

Table 18: **FE-verified improvement over the binary reference.** Columns specify required improvement 100γ; rates are percentages under the volume cap and obstacle constraint. **(a)** Five-candidate macro rates weight generator–split conditions equally; bold marks higher point estimates among the two rules that rank without FE labels. **(b)** RF16+CNO rates retain all planned instances per split; Macro weights splits equally.

Rule / split	0%	0.1%	1%	2%	5%
(a) Five-candidate pools: selector comparison					
Maximum volume	28.830	24.199	9.822	5.056	1.844
CNO	**28.946**	**25.009**	**11.435**	**6.387**	**2.939**
FE screening	31.681	26.757	12.012	6.718	3.000
(b) RF16+CNO: performance across splits					
Test	74.70	66.40	33.40	16.85	4.85
OOD-BC	50.20	45.50	23.10	11.70	3.40
Low-Re	78.40	69.60	31.60	17.20	4.40
Dual	60.00	52.67	23.33	14.67	8.00
Large	60.50	53.50	28.50	12.50	3.00
Notch	77.33	67.33	32.00	17.33	5.33
Macro	66.86	59.17	28.66	15.04	4.83

RF16+CNO selects reference-improving designs across all six splits. Table 18(b) evaluates this proposal-and-selection protocol, separately from the generator-averaged five-candidate comparison.

Most unselected RF16 proposals lack FE labels, so its rates describe selected designs, not a full-pool oracle. Figure 10 shows verified examples across all six splits.

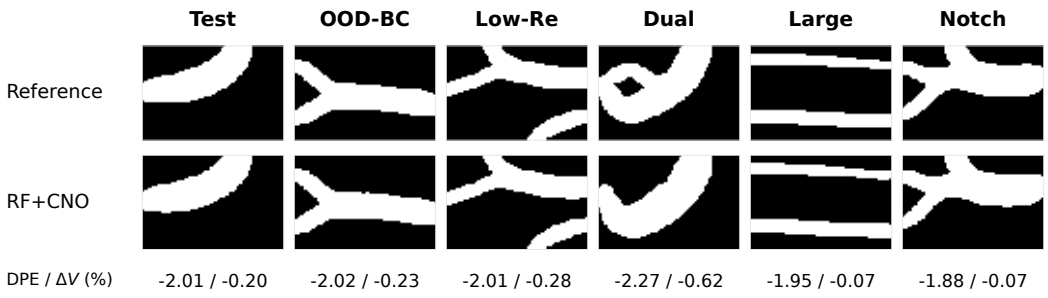


Figure 10: **Reference-improving RF designs selected by CNO.** Each column shares conditions; all selections satisfy the volume cap and fixed obstacles. Entries give DPE / signed volume deviation $\Delta V = 100(V/V_f - 1)$, both in percent. Examples illustrate diverse geometries near split-wise median improvements among designs improving by at least 1%; Table 18(b) reports improvement frequencies.

Candidate generation, predictor calls, and FE analyses are separate costs. Reference improvement uses binary-geometry objectives; continuous optimization endpoints are compared in Appendix H.

H Numerical Initialization and Optimization

H.1 Initialization Interfaces

All four interfaces receive the same floating-point generated density $\widehat{\rho}$ and use the same optimizer, continuation, and iteration stopping policy. Raw passes $\widehat{\rho}$ directly to the optimizer. The processed interfaces clip densities to $[0, 1]$, fix obstacles to solid and designated inlet cells to fluid, and leave outlets unforced.

Volume projection. Let A contain the free cells, F the inlet cells fixed to fluid, and N_e the total number of equal-area cells. All processed interfaces apply the same shift-and-clip projection to their intermediate density $\widetilde{\rho}_0$:

$$\rho_{0,e} = \mathrm{clip}_{[0,1]}(\widetilde{\rho}_{0,e} + s), \qquad \sum_{e \in A} \rho_{0,e} = \mathrm{clip}_{[0,|A|]}(V_f N_e - |F|). \tag{32}$$

The shift s enforces the remaining budget without changing fixed cells; ρ_{unif} is this budget divided by $|A|$. Projection-only adds no cleanup or relaxation.

Global and local relaxation. Global-relaxation blends free cells toward ρ_{unif}. SP-CPR retains inlet-connected components of the thresholded mask, or the largest nonempty component if none touches an inlet. It relaxes the cleaned soft field ρ_{clean} toward a signed-distance profile ρ_{sdf}, with spatial weights $w_{\mathrm{rel}} \in [0, 1]$ concentrated near interfaces and removed or locally inconsistent regions. Before projection, the two interfaces form

$$\widetilde{\rho}_0 = (1 - \lambda_{\mathrm{mix}})\widehat{\rho} + \lambda_{\mathrm{mix}}\rho_{\mathrm{unif}} \qquad \text{(Global-relaxation)}, \tag{33}$$

$$\widetilde{\rho}_0 = (1 - w_{\mathrm{rel}})\rho_{\mathrm{clean}} + w_{\mathrm{rel}}\rho_{\mathrm{sdf}} \qquad \text{(SP-CPR)}. \tag{34}$$

Both expressions apply only to free cells. The shared global weight $\lambda_{\mathrm{mix}} = 0.25$ is selected on RF validation instances, primarily by $U_P(0.001)$, the 0.1% relaxed endpoint criterion. Connectivity, relaxation, and projection settings are specified in the accompanying implementation.

H.2 Completion and Endpoint Quality

The joint rate $U_P(\epsilon)$ in Eq. (6) retains every scheduled attempt with its original weight. Completion means a finite normal return under the stopping policy in Appendix A.2, including iteration-budget termination. Negative ϵ requires an improvement over the continuous cold-start objective; zero

requires parity or better. Paired intervals resample physical instances within each split, keeping their generator and sampling records together (Appendix C.2).

Table 19: **Initialization interfaces change completion and endpoint quality.** Rates are percentages with all attempts retained; thresholds use the continuous cold objective. **(a)** Four-interface comparison, pooled with the observed attempt weights. **(b)** Raw/SP-CPR comparison within each generator. Bold marks the highest point estimate in each column within a panel.

(a) Four initialization interfaces

Interface	Completion	$U_P(-1\%)$	$U_P(-0.1\%)$	$U_P(0)$	$U_P(0.1\%)$
Raw	94.07	5.51	13.47	24.26	39.07
Projection-only	**99.85**	6.56	22.38	52.55	68.66
Global-relaxation	99.75	6.72	26.86	69.74	**94.80**
SP-CPR	99.67	**8.30**	**32.11**	**71.56**	92.18

(b) Generator-dependent effects

	$U_P(0)$		$U_P(0.1\%)$		$U_P(1\%)$	
Generator	Raw	SP-CPR	Raw	SP-CPR	Raw	SP-CPR
CNN	18.33	**77.20**	35.07	**95.98**	65.24	**98.38**
cGAN	20.89	73.07	42.64	94.58	78.96	97.91
DDPM	**50.18**	71.78	**71.00**	91.96	**88.04**	96.04
RF	20.93	75.96	33.27	94.67	56.93	97.27
EDM	16.60	60.38	24.58	84.22	47.80	93.24
PhysToken	18.60	70.96	27.89	91.69	51.02	95.20

Table 19(a) separates solver completion from endpoint quality. Projection-only has the highest completion rate, while global and local relaxation substantially increase cold-start parity. SP-CPR leads at the reported non-positive thresholds; Global-relaxation leads at the relaxed criterion used to select its mixing weight. Panel (b) shows that SP-CPR improves all six generators at every listed threshold. The highest parity rate changes from DDPM under Raw to CNN under SP-CPR, showing that generator comparisons depend on the initialization interface.

The split-wise contrasts in Figure 6(c) preserve this threshold dependence. At $\epsilon = -0.001$, every split's pointwise paired 95% interval favors SP-CPR over Global-relaxation. At $\epsilon = 0.001$, all point estimates favor Global-relaxation, but the Notch interval includes zero. The accompanying results retain the full split-wise intervals.

H.3 Paired Improvements and Computational Scope

Objective improvement. Of the 47.30-percentage-point increase in $U_P(0)$, 43.19 points come from quality transitions among attempts completed under both Raw and SP-CPR. In that common population, SP-CPR lowers the objective in 84.25% of attempts, with a median reduction of 0.327 percentage points relative to the matched cold objective. The rate gain and median reduction measure different aspects of endpoint improvement.

Same density, different endpoints. Figure 11 traces a shared generated density through Raw, Global-relaxation, and SP-CPR. In the Test double-pipe example, SP-CPR reaches -0.89% relative to the common cold objective and passes the strict -0.1% target, whereas Global-relaxation reaches $+0.05\%$. In the OOD-BC bend, Global-relaxation reaches -0.30% and passes the relaxed $+0.1\%$ target, while SP-CPR reaches $+0.16\%$. The cases illustrate the threshold-dependent comparison in Figure 6; they are selected contrasts rather than estimates of its frequency.

Optimization work. Relative to the recorded cold runs, SP-CPR reduces outer iterations by 10.1–27.5% across generator–split conditions. The recorded mean solver-stage time reduction is 28.7%, including preprocessing and optimization but excluding initial-response acquisition and generator inference. Cold times come from separate data-generation runs; launch paths and timeout

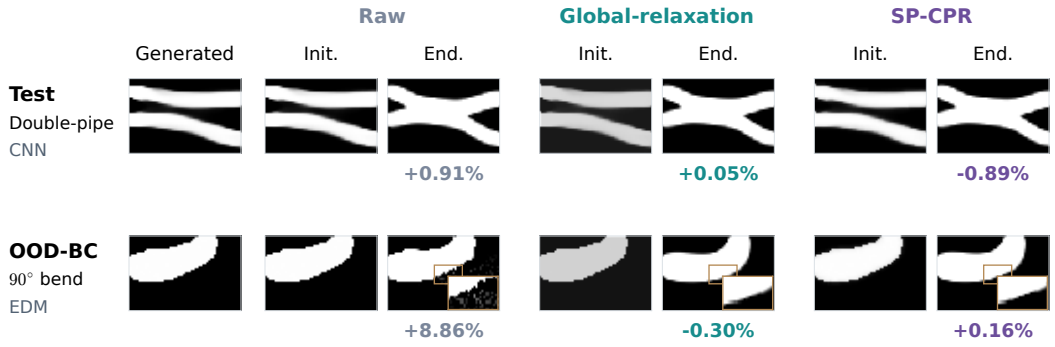


Figure 11: **Identical generated densities, different initialization interfaces.** Each example shares the generated input, physical conditions, optimizer, stopping policy, and cold-start comparator across all three branches. Initial and terminal densities use the same grayscale range $[0, 1]$. Endpoint labels give $100(J_{\text{end}}/J_{\text{cold}} - 1)$. The Test double-pipe illustrates SP-CPR's strict-quality benefit; the OOD-BC 90° bend illustrates a relaxed-threshold trade-off. Embedded bend insets enlarge the same cell window in each endpoint. These cases were selected for clear structural changes and opposite threshold outcomes. Isolated visualization replays match the historical objectives and iteration counts.

accounting also differ between the four interfaces. These timings describe solver-stage work, not a hardware-matched four-way or end-to-end speed comparison.

Selection before refinement. On a separate 96-instance confirmation set, fixed-candidate, CNO, and maximum-volume selection use the same candidate population. Selected floating-point densities initialize the optimizer. With SP-CPR, fixed-candidate and CNO selection both attain $U_P(0) = 75.00\%$, and CNO has no higher point estimate at the other thresholds. The set shows no added endpoint gain from CNO selection; equal rates do not establish equivalence. Full results accompany the evaluation records.