METAGROUND: PHYSICS-GROUNDED CONDITIONAL DIFFUSION FOR NONLINEAR METAMATERIAL INVERSE DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models for nonlinear metamaterial inverse design are typically confined to a single mechanical property, overlooking the physical coupling between the stress–strain (SS) response and volume fraction (VF). These generators rely on surrogate proxies to rapidly pre-screen candidate geometries, yet evaluating with proxies not validated against high-fidelity FEM ground truth produces an optimistic bias that inflates apparent performance, and low-accuracy pre-screening in turn requires more inference candidates for compensation.

We present **MetaGround**, a DBTL-style framework in which conditional diffusion designs SDF candidate geometries, FEM simulation provides high-fidelity FEM ground truth, a surrogate proxy tests candidate quality, and residual-guided adaptation learns to recalibrate the proxy. Within this loop, MetaGround introduces multi-physics conditioning architectures for SS and VF, a VF extrapolation protocol probing the generalization boundary, FEM-supervised adaptation driven by residual-guided selection, and finally an independent test set sampled from the full design space to anchor evaluation in FEM ground truth. Experiments show that the cross-attention + FiLM configuration (B2-PD) outperforms the baseline and is the most robust under deep VF extrapolation; the independent test set systematically confirms the optimistic bias of the proxy, and residual-guided selection improves proxy accuracy, raising the downstream screening pass rate from 34.4% to 39.6% at a fixed FEM budget.

1 INTRODUCTION

Mechanical metamaterials derive their effective behavior from architecture. A single periodic unit cell can be designed to exhibit a prescribed nonlinear stress–strain (SS) response while using substantially less material than a dense solid, rendering inverse design compelling for lightweight structures and energy absorption. The inverse mapping is one-to-many: distinct geometries can manifest comparable macroscopic responses, so we need a generative model that yields candidate distributions rather than a single solution.

Conditional diffusion models are well suited for this formulation (Liu et al., 2025; Zheng et al., 2026). A signed distance function (SDF) provides a continuous geometry representation, while a U-Net denoising diffusion probabilistic model can sample SDF realizations conditioned on a prescribed response. Most existing generators, however, condition on a single mechanical property or treat additional properties as independent scalar labels (Bastek & Kochmann, 2023; Wang et al., 2024). In practice, the SS response and volume fraction (VF) are physically coupled: reducing VF alters stiffness, load-transfer pathways, and contact regions, so joint SS+VF conditioning is not a conventional multi-label problem. Furthermore, the surrogate proxies used to rank generated candidates are trained on the same data distribution as the generators, without independent validation against high-fidelity FEM ground truth, creating an evaluation optimism that inflates apparent performance.

This work addresses three interrelated questions: (i) Can a conditional diffusion model jointly conditioned on SS and VF generate geometries whose realized VF conforms to the prescribed target while preserving SS fidelity, and how do conditioning architectures compare? (ii) At what VF ex-

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.

trapolation magnitude does the practical generalization boundary arise when the generator is pushed beyond its training distribution? (iii) Can FEM-derived ground-truth labels adapt a pretrained proxy to improve its predictive fidelity?

Contributions.

- A systematic comparison of three multi-physics conditioning architectures (concat, cross-attention + FiLM, concat + FiLM) across 12 generator arms, establishing cross-attention + FiLM as the best single configuration in our protocol for joint SS+VF control and identifying its per-physics-dropout variant (B2-PD) as the most robust under deep VF extrapolation.
- A volume-fraction extrapolation protocol that probes the generalization boundary of SDF-based nonlinear metamaterial diffusion via a systematic six-level FEM campaign.
- A Design–Build–Test–Learn (DBTL) framework that closes the loop between conditional generation, proxy ranking, FEM validation, and FEM-supervised proxy adaptation, in which residual-guided selection mitigates evaluation optimism and an independent test set exposes the optimistic bias of proxy assessment.

2 Related Work

Physics-constrained generative metamaterial design. Physics-constrained generative models incorporate physical structure into the generative process; within this direction, Liu et al. (2025) employ an SDF-based U-Net diffusion generator conditioned on a single stress–strain target. We target three open aspects: joint SS+VF conditioning across three compared architectures, a VF extrapolation protocol probing generalization beyond the training distribution, and FEM-supervised proxy adaptation with independent validation. The closest competitor, DiffuMeta (Zheng et al., 2026), reports extrapolation beyond the training domain but fixes relative density at $\rho{=}0.1$, does not treat VF as a controllable coupling target, and neither probes the boundary systematically nor adapts its surrogate; our contribution is positioned by capability completion rather than same-benchmark comparison. Bastek & Kochmann (2023) encode SS curves as temporal sequences, later extended to multi-material design (Park et al., 2026); we employ SDF spatial representation with VF as a coupled condition. Related approaches span unified generation–prediction models, density-field and graph-space designers, physics- and manufacturing-aware generation, spinodoid design, hybrid neural representations, and human–AI collaboration (Zhan et al., 2025; Yang et al., 2024; Maurizi et al., 2025; Li et al., 2026a; Xie et al., 2026; Li et al., 2026b; Rosenkranz et al., 2025; Xue et al., 2025; Qi et al., 2024), and multi-objective diffusion for auxetic patches that steers an unconditional model via surrogate guidance (Chen et al., 2026), building on the SDF representation of DeepSDF (Park et al., 2019) and surveyed in (Long et al., 2024). Extrapolation has been quantified mainly for discrete representations (Polat et al., 2026; Antoniuk et al., 2026), whereas continuous-field generative models with mechanics-coupled conditions remain largely unprobed; our protocol (r_0–r_{30}) provides a systematic VF-extrapolation study in SDF-based metamaterial diffusion, anchored by an independent FEM campaign.

Conditioning architectures. Diffusion models and their variants (DDPM (Ho et al., 2020), DDIM (Song et al., 2021a), score-based SDEs (Song et al., 2021b), latent diffusion (Rombach et al., 2022), and diffusion transformers (Peebles & Xie, 2023)) provide the generative foundation; classifier-free guidance (Ho & Salimans, 2021) is our conditional-sampling strategy, FiLM (Perez et al., 2018) offers lightweight feature-wise modulation, and surveys (Cao et al., 2026) with Pareto-conditioned diffusion (Shrestha et al., 2026) delineate the multi-property landscape. In the metamaterial setting, existing generators typically inject a single heterogeneous condition (an SS curve) via plain concatenation; which conditioning architecture best executes *joint* SS+VF conditioning remains unresolved. Our benchmark (concat, cross-attention+FiLM, concat+FiLM) across generator arms fills this gap.

Surrogate models and proxy adaptation. Forward surrogates, from neural operators such as the Fourier Neural Operator (Li et al., 2021) and DeepONet (Lu et al., 2021) to MLP-based and hybrid architectures (Jin et al., 2025; Upadhyay et al., 2024), learn PDE solution mappings; in metamaterial design, NOT (Liu et al., 2025) predicts SS responses from SDF geometries for candidate ranking. Yet such surrogates are rarely validated against independent high-fidelity FEM ground truth:

prior work stops at proxy-level ranking or validates candidate quality without isolating proxy error. Tauberschmidt et al. (2026) enforce physical consistency by fine-tuning flow-matching generative models through differentiable PDE residual minimization, supervising the generative model with continuous residuals, whereas ours targets a separate forward surrogate with discrete FEM labels. Yang et al. (2025) likewise observe that out-of-distribution surrogate errors can produce adversarial designs yet counter them at the generation level through a generative prior, whereas we adapt the proxy directly with FEM ground truth and validate the gains on an independent test set. Closed-loop workflows (Babu et al., 2026; Ito et al., 2026) fit surrogates from scratch or target other domains; none adapts a pre-trained proxy with FEM labels or validates gains on independent FEM test sets. We adapt an existing predictor via two trainers (frozen adapter, fine-tuning), extended by residual-guided selection with discrete FEM labels. Finally, our proxy acts as a reward model over generated candidates; its in-distribution optimism is a physical analogue of reward overoptimization (Gao et al., 2023), with FEM ground truth and DBTL feedback as the counterparts of preference data and online correction.

3 METHOD

MetaGround closes a Design–Build–Test–Learn (DBTL) feedback cycle (Figure 1) comprising a multi-physics conditional diffusion generator, a surrogate ranking proxy, and proxy adaptation supervised by finite-element-method (FEM) ground truth and driven by residual-guided selection. In DBTL terms, the generator designs candidate geometries conditioned on a target stress–strain curve and a prescribed VF extrapolation ratio r; FEM simulation builds high-fidelity ground-truth labels for the proxy-ranked candidates; an independent test set tests the proxy against that ground truth, exposing its optimistic bias; and residual-guided selection learns by recalibrating the proxy on the FEM-labeled pool, whereas at evaluation time the FEM branch is evaluation-only. We accordingly address three interrelated problems: (i) how to jointly condition on the physically coupled SS and VF targets; (ii) where the practical generalization boundary of VF extrapolation lies; and (iii) how to recalibrate the ranking proxy with FEM ground-truth labels, sampling where the proxy residual is largest, to improve its predictive fidelity.

3.1 PROBLEM FORMULATION

A 2D periodic unit cell is represented as a 120×120 signed distance function (SDF) grid on $[0, 1]^2$, where the solid phase is $\{\text{SDF} < 0\}$ and the void phase is $\{\text{SDF} > 0\}$. The volume fraction is

$$\text{VF} = \mathcal{V}(\text{SDF}), \qquad \mathcal{V}(s) = \langle \mathbb{I}[s < 0] \rangle, \tag{1}$$

A candidate is deemed geometrically valid if it preserves the unit-cell periodicity and contains no island (an isolated solid region disconnected from its periodic images). A neural operator transformer (NOT) $\mathcal{N}_\theta$ maps an SDF grid to a 51-point SS curve under plane-strain compression; the forward error is

$$E_{\text{NOT-FEM}} = \frac{\|\mathcal{N}_\theta(\text{SDF}) - \text{SS}_{\text{FEM}}\|_2}{\|\text{SS}_{\text{FEM}}\|_2}. \tag{2}$$

A multi-physics conditional diffusion model G_ϕ samples SDFs given $(\text{SS}_{\text{target}}, \text{VF}_{\text{target}})$; the inverse error is

$$E_{\text{target-FEM}} = \frac{\|\text{SS}_{\text{target}} - \text{SS}_{\text{FEM}}(\text{SDF})\|_2}{\|\text{SS}_{\text{target}}\|_2}. \tag{3}$$

We also define the proxy-only ranking error

$$E_{\text{target-NOT}} = \frac{\|\text{SS}_{\text{target}} - \mathcal{N}_\theta(\text{SDF})\|_2}{\|\text{SS}_{\text{target}}\|_2}, \tag{4}$$

used for initial screening before FEM validation. Proxy adaptation minimizes $E_{\text{NOT-FEM}}$ on unseen geometries via FEM-supervised calibration. Given a source geometry with $\text{VF}_{\text{source}}$, the target VF is set to

$$\text{VF}_{\text{target}} = \text{VF}_{\text{source}} \times (1 - r), \quad r \in \{0,\, 0.10,\, 0.15,\, 0.20,\, 0.25,\, 0.30\}. \tag{5}$$

The protocol provides six discrete probes of each model's VF extrapolation capability. The achieved VF extrapolation ratio is

$$r_{\text{ach}} = \frac{\text{VF}_{\text{source}} - \text{VF}_{\text{achieved}}}{\text{VF}_{\text{source}}}, \tag{6}$$

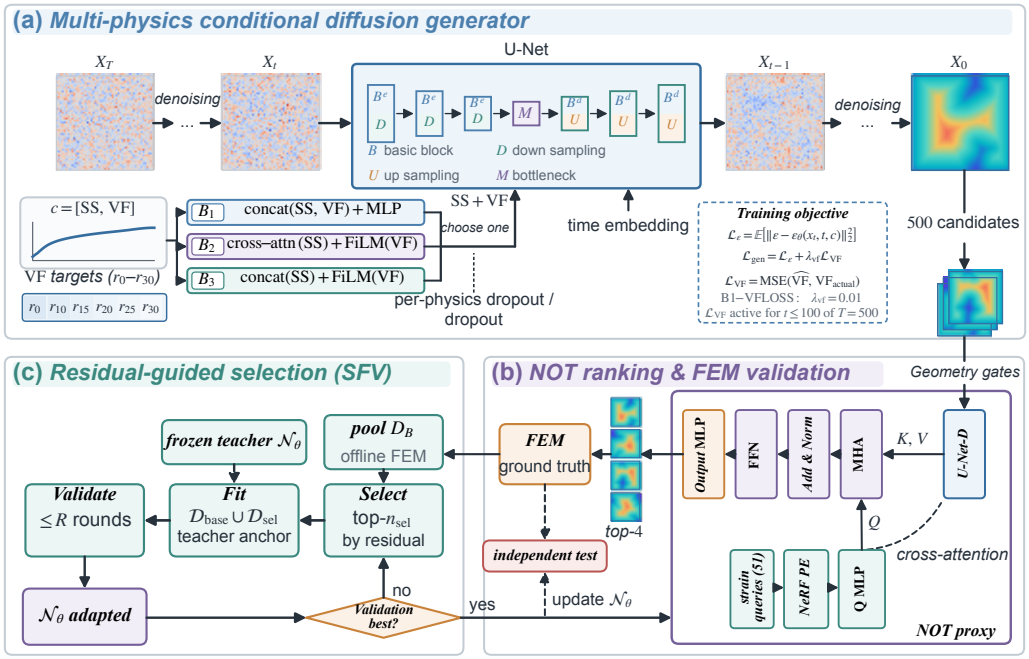


Figure 1: MetaGround system overview. **(a) Multi-physics conditional diffusion generator.** A target SS curve and a requested VF target form the conditioning vector $\mathbf{c}$, injected into the U-Net through one of three parallel conditioning architectures (B1 / B2 / B3). **(b) NOT ranking and FEM validation.** The generator produces 500 candidate SDFs per (arm, target, level) triple; periodicity and no-island gates reject geometrically invalid candidates; the NOT proxy $\mathcal{N}_\theta$ ranks survivors by proxy SS fidelity; the top-4 candidates are validated against Abaqus FEM ground truth. **(c) Residual-guided selection (SFV) loop.** Select: select the top-n_{sel} pool samples by proxy residual; Fit: fine-tune the proxy on $\mathcal{D}_{\text{base}}^{\text{train}} \cup \mathcal{D}_{\text{sel}}^{(r)}$ with the frozen teacher $\mathcal{N}_{\theta_T}$ anchoring the ℓ_2 objective; Validate: keep the val-best checkpoint under a budget of at most R rounds; the returned $\mathcal{N}_{\theta^*}$ replaces the ranking proxy in (b). See Algorithm 1 for the detailed workflow.

where $\text{VF}_{\text{achieved}}$ is the VF of the FEM-validated generated geometry. Whether a target is counted as reached is decided by the pass criteria in Appendix A.4. Because G_ϕ and $\mathcal{N}_\theta$ are trained on the same source distribution, $E_{\text{NOT-FEM}}$ on model-generated geometries understates the true error over the full design space; Section 4.3 diagnoses this optimism and Section 3.3 mitigates it with FEM labels.

3.2 Multi-Physics Conditional Diffusion Generator

We adopt a denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) operating on 120×120 SDF grids. The forward process perturbs a pristine SDF $\mathbf{x}_0$ over T timesteps, sampled via reparameterization

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The reverse process reconstructs $\mathbf{x}_0$ level by level through the Gaussian posterior $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, whose mean is given by the noise estimator $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}) \right), \tag{8}$$

The estimator is trained with the standard ℓ_2 noise-prediction loss:

$$L_\epsilon = \mathbb{E}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2 \right]. \tag{9}$$

The conditioning vector $\mathbf{c}$ encodes two physical objectives: a 51-point SS curve and a scalar VF. Sampling uses classifier-free guidance: during training $\mathbf{c}$ is set to empty with probability p_{uncond}; at

inference the conditional and unconditional estimates are combined with guidance weight w,

$$\tilde{\boldsymbol{\epsilon}}_\theta = (1 + w)\, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}) - w\, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \emptyset), \tag{10}$$

followed by ancestral sampling over T steps. We compare three conditioning mechanisms (expanded in Figure 1a). The U-Net backbone is specified in Appendix B.1.

B1 (plain concatenation). The SS curve (51-dim) and the VF scalar are concatenated into a conditioning vector and mapped through a shared 4-layer MLP into a single conditioning embedding. A single VF dimension would occupy only $\sim 2\%$ of the 52-dim concatenation, so we repeat it to 16, giving the VF a reasonable share of the conditioning vector (16/67 $\approx$ 24%, rather than $1/52 \approx 2\%$), at negligible additional parameters. This $\times 16$ repetition is the uniform default across all VF-conditioned arms, so B1, B2, and B3 share an identical conditioning vector.

B2 (cross-attention + FiLM). The SS curve is tokenized into a sequence and injected via cross-attention layers at the bottleneck, endowing SS with an independent encoding pathway. The VF scalar is injected via Feature-wise Linear Modulation (FiLM): at each residual block, VF produces a per-channel scale γ and shift β that modulate the GroupNorm output. SS, VF, and time thus follow separate injection pathways (attention, multiplicative gating, additive embedding). **B3 (concat + FiLM).** SS is injected via plain concatenation (as in B1), while VF is injected via FiLM modulation (as in B2). All three variants share the DDPM U-Net backbone of Appendix B.1 (self-attention at levels 3–4).

Auxiliary VF loss. When VF is injected via plain concatenation (B1), the network may partially disregard the VF dimension. To promote VF sensitivity, an auxiliary VF loss penalizes the discrepancy between the VF implied by the model's denoised prediction and the ground-truth VF. At each training step, for timesteps $t \le t_{\text{max}}$ and temperature τ, the soft VF and the actual VF are

$$\hat{\text{VF}} = \mathcal{V}_\tau(\hat{x}_0), \qquad \text{VF}_{\text{actual}} = \mathcal{V}(x_0), \tag{11}$$

and the auxiliary VF loss

$$L_{\text{VF}} = \text{MSE}(\hat{\text{VF}},\ \text{VF}_{\text{actual}}) \tag{12}$$

is added to the total objective with weight λ_{vf} (values and ablation in Appendix B.3).

Architecture variants. Each backbone has dropout and auxiliary-loss variants. The **-D** variant applies network-level dropout ($p = 0.1$) on the conditioning embedding. The **-PD** (per-physics dropout) variant independently drops the SS and VF conditioning channels with probability $p = 0.1$. The **-VFLOSS** variant adds the auxiliary VF loss described above.

3.3 NOT Proxy and FEM-Supervised Adaptation

NOT (Liu et al., 2025) is an encoder–decoder Transformer mapping SDF grids to 51-point SS curves, serving as a fast surrogate for Abaqus FEM simulation during candidate ranking; architecture details in Appendix B.2 (Figure 3). The optimistic-bias hypothesis is tested against two FEM-anchored test sets, the in-domain $\mathcal{D}_{\text{indom}}$ and an out-of-domain set generated independently from the full design space; their construction and scale are given in Appendix A.1. Adaptation must serve two goals at once: improving accuracy within the FEM-labeled distribution while retaining cross-distribution robustness. We therefore propose two trainers (defined in Appendix D.1): S1, a frozen-backbone residual adapter, which leaves the backbone untouched and appends a lightweight residual head after the decoder, so that only the residual head is updated; and S2, fine-tuning with a pure MSE objective, which unfreezes the backbone and updates all parameters on the FEM-labeled distribution.

Residual-guided selection loop. Residual-guided selection requires the proxy to absorb newly labeled samples and re-estimate residuals quickly each round, a capacity that depends on the learning capacity of fine-tuning; we therefore extend S2 into a closed-loop procedure, residual-guided selection: each round selects the pool samples with the largest current proxy residual and fine-tunes the proxy on them, in three phases: Select, Fit, and Validate (SFV). The SFV loop is therefore an extension of the fine-tuning trainer rather than a separate strategy, which we ablate in Section 4.3. The full procedure is given in Algorithm 1 (Appendix D.2).

Each round selects n_{sel} samples from the $\mathcal{D}_{\text{indom}}$ pool. The training objective makes the anchoring scope explicit:

$$\mathcal{L} = \underbrace{\|\mu_s - y\|^2}_{\text{all training rows}} + \lambda \underbrace{\|\mu_s - \mu_T\|^2}_{\mathcal{D}_{\text{base}}^{\text{train}} \text{ rows only}}, \tag{13}$$

where y is the ground-truth target: the SS label for $\mathcal{D}_{\text{base}}^{\text{train}}$ rows and the FEM curve for newly selected $\mathcal{D}_{\text{indom}}$ rows and the teacher anchor applies only to $\mathcal{D}_{\text{base}}^{\text{train}}$ rows. Here "teacher" denotes regularization, not distillation: the anchor applies only to $\mathcal{D}_{\text{base}}^{\text{train}}$ rows, preventing catastrophic forgetting in the spirit of replay-with-distillation (Li & Hoiem, 2016; Rebuffi et al., 2017), while the primary supervision is FEM ground truth on newly labeled samples.

4 Experiments

4.1 Setup

The SDF dataset (Liu et al., 2025) provides 10,000 unit-cell geometries with Abaqus-verified SS curves, split $8{,}000/2{,}000$ into a training split $\mathcal{D}_{\text{base}}^{\text{train}}$ and a validation split $\mathcal{D}_{\text{base}}^{\text{val}}$; we additionally generate 1,000 geometries independently from the full design space to form a test split $\mathcal{D}_{\text{base}}^{\text{test}}$. All generator variants are trained on $\mathcal{D}_{\text{base}}^{\text{train}}$ and checkpoint-selected on $\mathcal{D}_{\text{base}}^{\text{val}}$. Every arm is evaluated on an in-domain dataset $\mathcal{D}_{\text{indom}}$ and an out-of-domain dataset $\mathcal{D}_{\text{indep}}$, both selected by VF-stratified SS targets and both driven through a unified inference and simulation protocol (Appendix A.4) to produce candidates and obtain FEM ground truth; per-target composition, splits, and provenance are given in Appendix A.1. We report the per-case pass rate, the task-level pass@4, and the proxy error $E_{\text{NOT-FEM}}$ across the six VF extrapolation levels r_0–r_{30}; pass criteria and denominator conventions are given in Appendix A.4, and the evaluation pipeline in Section 3 and Figure 1.

4.2 Multi-Physics Conditioning and VF Extrapolation

Three conditioning mechanisms are compared: B1 concatenates the SS and VF channels, B2 separates modalities via cross-attention on SS and FiLM on VF, and B3 concatenates SS while modulating VF through FiLM. Each mechanism is tested in three variants: base, network dropout, and per-physics dropout, denoted -D and -PD respectively. All arms are evaluated against an SS-only baseline B0 (Liu et al., 2025) under the same protocol: the per-case pass rate under an inverse error $E_{\text{target-FEM}} \leq 5\%$ plus an asymmetric VF tolerance ϵ is reported in Table 1 across six VF extrapolation levels r_0–r_{30}, and task-level performance as pass@4 in Table 13. In-domain evaluation on $\mathcal{D}_{\text{indom}}$ uses the 18 VF-stratified targets and is reported with isomorphic matrices in Appendix Tables 8 and 9. Every conclusion below first reports the in-domain result and then tests its generalization on the independent set.

Table 1: $\mathcal{D}_{\text{indep}}$: per-case pass rate (%), 44 cases per cell (11 targets × top-4); B0 exists only at r_0.

Family	Model	r_0	r_{10}	r_{15}	r_{20}	r_{25}	r_{30}	ALL
SS-only	B0	43.2	—	—	—	—	—	—
B1 (concat)	B1	34.1	45.5	43.2	36.4	25.0	20.5	34.1
	B1-D	27.3	34.1	34.1	47.7	31.8	22.7	33.0
	B1-PD	43.2	31.8	38.6	40.9	34.1	25.0	35.6
B2 (xattn+FiLM)	**B2**	**50.0**	43.2	**59.1**	43.2	**40.9**	22.7	**43.2**
	B2-D	34.1	36.4	36.4	47.7	38.6	25.0	36.4
	B2-PD	47.7	43.2	43.2	40.9	38.6	**34.1**	41.3
B3 (concat+FiLM)	B3	20.5	4.5	9.1	18.2	13.6	6.8	12.1
	B3-D	25.0	20.5	22.7	18.2	9.1	11.4	17.8
	B3-PD	43.2	22.7	20.5	18.2	9.1	6.8	20.1

Comparison against the SS-only baseline. B1 shares B0's UNet backbone, simply appending VF16. At in-domain r_0, B0 reaches 56.9% while B1 and B2 drop to 41.7% and 37.5%; at r_0 on

the independent set, B0 and B1 fall to 43.2% and 34.1%, whereas B2 rises to 50.0%, Tables 1 and 8. The two domains agree on B1, whose concatenated VF degrades SS reproduction, but disagree on B2, which falls below the baseline in-domain yet exceeds it out-of-domain, showing that the VF condition is not a free gain and its value depends on the injection pathway. We attribute this to the baseline's training memorization: at in-domain r_0, where the criterion reduces to SS-only reproduction with the VF tolerance waived, B0's median $E_{\text{target-FEM}}$ of 0.043 is the only one below the 5% threshold, versus 0.066 for B2; once distribution shifts, memorization fails and B2's conditioning-based generalization takes over. B1's shared concatenation mixes the 16-dim VF signal into the same MLP as the 51-dim SS signal, biasing capacity toward the higher-dimensional SS and contaminating the SS path, explaining its drop below B0 at r_0 in both domains; B2 isolates VF through FiLM, whose r_0 modulation acts as a material-quantity constraint consistent with the source geometry, improving SS reproduction.

B2 (cross-attention + FiLM) leads the conditioning architectures overall. The per-case ordering B2 $>$ B1 $>$ B3 holds in-domain, with ALL rates of 41.9%, 39.8%, and 20.4%, and is preserved on the independent set, 43.2%, 34.1%, and 12.1%, Tables 8 and 1; at the task level, in-domain B2 ties B1 at 73.1%, while on the independent set B2 ties B2-PD and B1-VFLOSS-w01 at 74.2%, Tables 9 and 13. The ordering survives distribution shift, and B2's relative edge becomes more pronounced out-of-domain. The in-domain task-level tie arises because in-distribution SS curves compensate the concat family's architectural deficit through memorization. Per-level trajectories for both metrics are in Appendix Figures 4 and 5. B2 also leads on efficiency: 2.326 M parameters versus 4.190 M for B1, and roughly 159 s to generate 500 candidates versus 204 s for B1 (Appendix B.5). B2's full modality separation keeps the SS cross-attention pathway free of VF contamination during feature computation, with FiLM applying a per-channel affine transform orthogonal to the time and SS signals. B1's shared MLP allocates capacity unevenly between the 51-dimensional SS and 16-dimensional VF, biasing toward the higher-dimensional SS. B3 retains B1's SS concatenation bottleneck while switching only the VF pathway to FiLM; since SS remains the dominant signal, this partial change cannot resolve plain concatenation's representational inadequacy.

Variant ablations. The dropout variants' gains are shift-specific. In-domain, both -PD and -D give essentially no overall gain, with pass@4 decreasing across the board, Appendix Table 9; on the independent set, -PD benefits the families without separation most, pass@4 rising by 9.1 pp for B3 and 4.5 pp for B1 and unchanged for the separated B2 at 74.2%, while -D lowers B1/B2 and raises only B3. In-distribution models are already well fit, so dropout offers no generalization benefit and only capacity loss. Explicit modality separation therefore matters precisely under distribution shift; per-dataset breakdowns are in Appendix B.4. The auxiliary VF loss shows the same shift-specific pattern. In-domain, B1-VFLOSS at $\lambda_{\text{vf}} = 0.01$ gains only +0.9 pp, from 39.8% to 40.7%, and turns slightly negative at r_{30}, from 33.3% to 30.6%, since in-domain VF obedience already suffices; on the independent set the gain is large, ALL rising from 34.1% to 42.0%, +7.9 pp, with the largest gain at r_{30}, 34.1% vs. 20.5%, where VF obedience matters most. Doubling the weight to $\lambda_{\text{vf}} = 0.05$ degrades sharply at deep extrapolation on the independent set; the independent-set comparison is in Table 12 and the in-domain counterpart in Appendix Table 6.

VF extrapolation. The volume-fraction sweep prescribes progressively lower VF targets relative to the source geometry, systematically probing the generator's generalization. In-domain, per-case pass rates decline markedly beyond roughly 20% reduction, the r_{25}–r_{30} regime, Appendix Table 11; the independent set shows the same trend, with the generator producing geometries with broken load paths or excessive compliance. At deeper reductions the prescribed VF moves further from the source operating point, and to remain geometrically valid the generator shifts its SS curve away from the target, the mean $E_{\text{target-FEM}}$ rising to 0.139 at r_{30}. The in-domain decline is milder: B2 holds 38.9% at r_{30} in-domain versus 22.7% on the independent set, as in-domain families carry larger mechanical margin under deep extrapolation. Qualitative zero-level-set geometries for a representative target are shown in Appendix Figure 6: at moderate reductions the generator reconfigures material within the extant structural archetype, mean $E_{\text{target-FEM}}$ 0.08–0.10. Fragility also varies by target: the effective boundary is governed by each geometry family's mechanical margin rather than by the preset ratio alone.

4.3 FEM-SUPERVISED PROXY ADAPTATION

This section first quantifies the optimism bias, then compares the two trainers along in-domain and out-of-domain dimensions, then examines the gains of the residual-guided loop under a fixed FEM budget, and finally localizes the mechanism through ablations and distills a compact recipe.

Proxy evaluation is optimistically biased. We evaluate the optimism bias on two test sets representing the in-domain and out-of-domain distributions: $\mathcal{D}^{\text{test}}_{\text{indom}}$ consists of model-generated candidates that lie within the NOT's training distribution, and $\mathcal{D}^{\text{test}}_{\text{base}}$ of 1,000 original OOD geometries representing the out-of-domain distribution. The NOT proxy achieves $E_{\text{NOT-FEM}} = 0.062$ and 0.174 on the two sets, Table 10. Generator outputs in $\mathcal{D}^{\text{test}}_{\text{indom}}$ remain topologically proximate to the NOT's training distribution even when conditioned on OOD target SS curves; the independent test set, drawn from the full design space with wider VF ranges and more diverse topologies, falls outside that distribution (full VF distributions in Table 3 and Figure 2). The proxy is reliable for ranking generator candidates yet degrades markedly on arbitrary OOD geometries, validating the necessity of FEM-grounded evaluation and motivating the FEM-supervised recalibration below.

S2 strategy. The S2 fine-tuning trainer adopts the closed-loop form of residual-guided selection. We ablate three dimensions: selection strategy, regularization strength λ, and per-round sample size; for selection strategy we compare random, non-cumulative, and cumulative residual selection, with trajectories in Figure 7 and hyperparameters and multi-seed results in Appendix D.2. Under a 10-round budget, cumulative residual selection reduces the out-of-domain evaluation error from 0.174 to 0.163 and the in-domain evaluation error from 0.062 to 0.049, improvements of $+6.4\%$ and $+20.4\%$, respectively; λ is insensitive between 0.5 and 2.0, with out-of-domain gains stable at $+6.4\%$ to $+6.5\%$; increasing the per-round sample size from 200 to 400 leaves the out-of-domain error unchanged while improving the in-domain gain from $+20.0\%$ to $+22.5\%$. The optimal configuration, 400 samples per round with $\lambda = 2.0$, is selected by validation loss and achieves $+6.5\%$ out-of-domain and $+22.5\%$ in-domain, with gains concentrating on hard samples: no hard error band degrades, and 46 of the 396 geometries with baseline error above 10% cross into $\leq 10\%$ (Figure 8). R_0, which applies the identical fine-tuning protocol to $\mathcal{D}^{\text{train}}_{\text{base}}$ only and introduces no $\mathcal{D}_{\text{indom}}$ samples, is statistically indistinguishable from the teacher baseline, showing that the gains require residual-guided accumulation of new $\mathcal{D}_{\text{indom}}$ samples rather than fine-tuning per se, Table 2.

The trade-off and mechanism of the two trainers. The S1 frozen-backbone adapter, trained on $\mathcal{D}^{\text{train}}_{\text{indom}}$, reduces the in-domain evaluation error from 0.062 to 0.057 ($+7.9\%$) and the out-of-domain evaluation error from 0.174 to 0.149 ($+14.3\%$); the S2 fine-tuning loop reduces the in-domain evaluation error from 0.062 to 0.048 ($+22.5\%$) and the out-of-domain evaluation error from 0.174 to 0.163 ($+6.5\%$), Table 2. The two trainers thus differ in where the mean error improves: S2 fine-tuning moves toward the adaptation distribution for in-domain accuracy, whereas S1's frozen backbone lowers the out-of-domain mean error further. That reduction must be read with the denominator-amplification caveat of Appendix A.4, and it does not confer a screening advantage (Appendix Table 16); we therefore report S1 as a mechanism control rather than as an out-of-domain-robust alternative. Ablations further localize the mechanism (Appendix D): gains are driven mainly by cumulative accumulation rather than residual ranking (a selection $\times$ accumulation study, Table 17), and a frozen-backbone adapter trained once on $\mathcal{D}^{\text{train}}_{\text{indom}}$ lowers the out-of-domain mean error further while giving up part of the in-domain gain, without a corresponding screening gain; replacing the loop's fine-tuning with an anchored adapter head collapses to the teacher, so the closed loop's in-domain gains require the fine-tuning + anchoring recipe.

Downstream screening utility. On $\mathcal{D}_{\text{indep}}$, we re-rank each cell's four candidates with the adapted proxy R_{10} versus the frozen SS-only NOT. The adapted proxy raises the top-1 FEM-valid rate from 34.4% to 39.6% ($N = 736$ cells), confirming that the proxy-accuracy improvement translates into downstream screening of generated candidates (per-cell results in Appendix Table 16; pool-size efficiency in Appendix D.3). The cumulative-versus-random comparison on the full 1,000-case test set is reported in Table 14 and Table 15. Taken together, these results define a compact recipe: start from the frozen teacher, adapt with the S2 fine-tuning trainer under teacher anchoring, and select hard samples cumulatively. The S1 adapter is cheaper and lowers the out-of-domain mean error further, but does not improve screening yield (Appendix Table 16); we report it as a mechanism control, not as a recommended trainer.

Table 2: Residual-guided selection results. Teacher: frozen NOT proxy. R_0: $\mathcal{D}^{\text{train}}_{\text{base}}$-only anchored fine-tuning. Cum-res: cumulative residual-guided selection. Improvements are relative to the teacher. $\mathcal{D}^{\text{test}}_{\text{base}}$ = independent 1K test set; $\mathcal{D}^{\text{test}}_{\text{indom}}$: the 463-case evaluation split of $\mathcal{D}_{\text{indom}}$. Trainer-ablation rows share the loop's pool and round budget. All adapted rows are 3-seed means $\pm$ std.

Configuration	$\mathcal{D}^{\text{test}}_{\text{base}}$ (1K)	$\mathcal{D}^{\text{test}}_{\text{indom}}$ (463)	Impr.
Teacher baseline	0.174	0.062	—
R_0 ($\mathcal{D}^{\text{train}}_{\text{base}}$ only)	0.174 ± 0.0014	0.062 ± 0.0001	—
Residual-guided selection			
Cum-res	$\mathbf{0.163 \pm 0.0005}$	0.048 ± 0.0003	$\mathbf{+6.5\%}$ / $+22.5\%$
Trainer ablation			
S1 adapter	$\mathbf{0.149 \pm 0.0004}$	0.057 ± 0.0002	$\mathbf{+14.3\%}$ / $+7.9\%$

5 Conclusion

We study conditional diffusion for SDF-based nonlinear metamaterial inverse design, unifying multi-physics conditioning, volume-fraction extrapolation, and proxy adaptation in one framework. Building on the DBTL paradigm, MetaGround spans the full inverse-design loop: multi-physics conditional generation, systematic probing of the physical-target extrapolation range, and FEM-anchored surrogate adaptation and validation. Extensive ablation over conditioning architectures and regularization variants identifies B2-PD as the best-matched multi-objective generator: it inherits the B2 family's strong multi-objective conditioning capability and, through per-physics dropout, extends the usable VF extrapolation range to deeper reductions. A systematic FEM campaign shows that B2-PD retains the highest per-case rate and pass@4 at the deepest extrapolation level r_{30}, 34.1% and 72.7%; per-case quality declines markedly beyond r_{25}, largely consistent with empirical failure under the current generator, ranking, and candidate budget. The independent test set reveals that proxy evaluation is optimistically biased, with $E_{\text{NOT-FEM}}$ of 0.174 versus 0.062; the S2 residual-guided loop improves in-domain evaluation by $+22.5\%$ and transfers to downstream screening ($34.4\% \rightarrow 39.6\%$), whereas the S1 frozen-backbone adapter, despite a larger reduction of the out-of-domain mean error, does not improve screening yield.

The principal implication is that trustworthy low-VF inverse design requires a closed feedback loop in which high-fidelity FEM ground truth continuously recalibrates the surrogate, rather than treating generation and evaluation as independent modules; the independent 11-target test set is consistent with architectural conclusions transferring to unseen geometries.

In future work, the framework can be applied to 3D SDF representations of nonlinear metamaterials; the strong multi-objective conditioning capability of B2 can be carried to other physical targets under multi-objective constraints; and the overall recipe can be transferred to other classes of materials. Combining the auxiliary VF loss with the B2 architecture remains a promising direction for further improvement.

AI Use Statement

We used generative AI tools for two purposes: to aid and polish the writing of this manuscript, including LaTeX formatting, and for retrieval and discovery of relevant literature, including reference organization and management. We did not use generative AI tools for research ideation or execution, for generating synthetic datasets, or for proving mathematical claims. We have reviewed all AI-assisted work: AI-generated text and suggestions were checked by the authors against the underlying experimental results, and the authors take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All quantitative claims in this paper are traceable to versioned runs and frozen analysis outputs. The conditional generator variants are specified in Section 3.2 with backbone configurations in Appendix B.1; the NOT proxy and the FEM-supervised adaptation procedures, including the residual-guided selection loop, are described in Section 3.3, with architecture details and multi-seed ablations in Appendices B.2 and D.2; evaluation metrics and pass criteria are defined in Appendix A.4; data splits and provenance, including the construction of the independent 1,000-geometry test set (Appendix A.2) and the exclusion of one anomalous target after a provenance audit, are documented in Appendix A.1; and the Abaqus configuration is specified in Appendix A.3. The proxy-adaptation experiments were run across three seeds, with the training and selection protocol given in Section 3.3 and Appendix D.2. Code is available at: `https://anonymous.4open.science/r/MetaGround-0418`; a de-anonymized repository will be released upon acceptance. The newly collected data, $\mathcal{D}_{\text{indom}}$ (4,608 generated geometries with FEM labels) and the independent test set, are available from the corresponding author upon reasonable request; the SDF dataset and the deployed NOT checkpoint originate from Liu et al. (2025) and remain subject to the terms of the original release.

REFERENCES

E. R. Antoniuk, P. Li, N. Keilbart, S. Weitzner, B. Kailkhura, and A. M. Hiszpanski. Active learning enables generation of molecules that advance the known pareto front. *npj Computational Materials*, 12(1):81, 2026. doi: 10.1038/s41524-025-01924-8.

A. Babu, R. A. Gouvêa, and G.-M. Rignanese. Towards automated discovery: A review of generative models, multimodal learning and closed-loop workflows in inverse materials design. *arXiv preprint arXiv:2606.02507*, 2026.

J.-H. Bastek and D. M. Kochmann. Inverse-design of nonlinear mechanical metamaterials via video denoising diffusion models. *Nature Machine Intelligence*, 5:1466–1475, 2023.

P. Cao, F. Zhou, Q. Song, and L. Yang. Controllable generation with text-to-image diffusion models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4771–4791, 2026.

Y. Chen, B. Graham, X. Mu, and S. Xiao. Diffusion-based inverse design of organ-inspired auxetic metastructure patches. *Frontiers in Materials*, 13:1885940, 2026. doi: 10.3389/fmats.2026.1885940.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

O. Ito, A. Katagiri, Y. Nakagawa, S. Saeki, J. Shiraishi, and M. Sasaki. Surrogate-assisted pedestrian protection design via a foundation model-orchestrated workflow. In *ICLR 2026 Workshop (FM4Science) Poster*, 2026.

Z. Jin, Z. Liu, N. Elabbasi, Z. Ulissi, G. X. Gu, and Z. Nie. Multi-physics inverse design of varifocal optical devices using data-driven surrogates and differential modeling. *arXiv preprint arXiv:2503.18911*, 2025.

H. Li, Y. Miao, M. Li, J. Bai, L. Wang, X. Liu, B. Gao, J. Liu, D. Mandic, Z. S. Khodaei, M. H. Aliabadi, and W. Chen. Steering topology distributions for unified generative design of architected metamaterials. *arXiv preprint arXiv:2607.24777*, 2026a.

J. Li, J. Guo, Y. Li, Z. Mao, J. Shen, T. Xu, D. Das, J. He, R. Hu, Y. Lee, K. Tsuda, and J. Shiomi. Inverse design of metamaterials with manufacturing-guiding spectrum-to-structure conditional diffusion model. *Cell Reports Physical Science*, 2026b. doi: 10.1016/j.xcrp.2026.103174.

Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, pp. 614–629, 2016.

Q. Liu, S. Koric, D. Abueidda, H. Meidani, and P. Geubelle. Toward signed distance function based metamaterial design: Neural operator transformer for forward prediction and diffusion model for inverse design. *Computer Methods in Applied Mechanics and Engineering*, 446:118316, 2025.

T. Long, Y. Zhang, and H. Zhang. Generative deep learning for the inverse design of materials. *arXiv preprint arXiv:2409.19124*, 2024.

L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021.

M. Maurizi, D. Xu, Y.-T. Wang, D. Yao, D. Hahn, M. Oudich, A. Satpati, M. Bauchy, W. Wang, Y. Sun, Y. Jing, and X. R. Zheng. Designing metamaterials with programmable nonlinear responses and geometric constraints in graph space. *Nature Machine Intelligence*, 7:1023–1036, 2025.

J. Park, S. Kushwaha, J. He, S. Koric, Q. Liu, I. Jasiuk, and D. Abueidda. Nonlinear inverse design of mechanical multi-material metamaterials enabled by video denoising diffusion and structure identifier. *Engineering Applications of Artificial Intelligence*, 172:114368, 2026.

J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

W. Peebles and S. Xie. Scalable diffusion models with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

C. Polat, E. Serpedin, and M. Kurban. How far can you grow? characterizing the extrapolation frontier of graph generative models for materials science. In *KDD '26*, pp. 11821–11832, 2026. doi: 10.1145/3770855.3818872.

J. Qi, Z. Jia, M. Liu, W. Zhan, J. Zhang, X. Wen, J. Gan, J. Chen, Q. Liu, and D. Ma. MetaScientist: A human-AI synergistic framework for automated mechanical metamaterial design. *arXiv preprint arXiv:2412.16270*, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, 2017.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

M. Rosenkranz, M. Kästner, and I. F. Sbalzarini. Data-efficient inverse design of spinodoid metamaterials. *Integrating Materials and Manufacturing Innovation*, 14:695–713, 2025.

J. Shrestha, S. Heiskanen, K. Hepola, S. Rissanen, P. Jääskeläinen, and J. Pajarinen. Pareto-conditioned diffusion models for offline multi-objective optimization. In *International Conference on Learning Representations (ICLR)*, 2026.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

J. Tauberschmidt, S. Fellenz, S. J. Vollmer, and A. B. Duncan. Physics-constrained fine-tuning of flow-matching models for generation and inverse problems. In *International Conference on Learning Representations*, 2026.

K. Upadhyay, J. N. Fuhg, N. Bouklas, and K. T. Ramesh. Physics-informed data-driven discovery of constitutive models with application to strain-rate-sensitive soft materials. *Computational Mechanics*, 2024.

H. Wang, Z. Du, F. Feng, Z. Kang, S. Tang, and X. Guo. Diffmat: Data-driven inverse design of energy-absorbing metamaterials using diffusion model. *Computer Methods in Applied Mechanics and Engineering*, 2024.

Z. Xie, W. Xu, D. Zhao, W. Zhang, D. Dong, B. Xu, N. Liu, S. Mao, and T. Xue. Physics-guided diffusion models for inverse design of disordered metamaterials. *npj Computational Materials*, 2026. doi: 10.1038/s41524-026-02328-y.

T. Xue, L. Liu, L. Lu, P. Henderson, P. Tang, H. Li, J. Liu, H. Zhao, H. Peng, and B. Bickel. MIND: Microstructure INverse design with generative hybrid neural representation. In *ACM SIGGRAPH 2025 Conference Papers*, 2025.

H. Yang, K. Azizzadenesheli, and H. Ren. Deep generative prior for first order inverse optimization. *arXiv preprint arXiv:2504.20278*, 2025.

Y. Yang, L. Wang, X. Zhai, K. Chen, W. Wu, Y. Zhao, L. Liu, and X.-M. Fu. Guided diffusion for fast inverse design of density-based mechanical metamaterials. *arXiv preprint arXiv:2401.13570*, 2024.

W. Zhan, J. Chen, D. Fu, and D. Zhou. Unimate: A unified model for mechanical metamaterial generation, property prediction, and condition confirmation. In *International Conference on Machine Learning*, pp. 74406–74421, 2025.

L. Zheng, S. Kumar, and D. M. Kochmann. Algebraic language models for inverse design of metamaterials via diffusion transformers. *Nature Machine Intelligence*, 8(4):628–640, 2026.

A Data and Evaluation Protocol

A.1 Data Splits and Provenance

Constructing $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{indom}}$. The base data domain $\mathcal{D}_{\text{base}}$ consists of the 10,000 unit-cell geometries of the original SDF dataset (Liu et al., 2025) (split group-disjointly 8:2 into 8,000 training and 2,000 validation geometries) plus the 1,000 test geometries we generate independently. The in-domain dataset $\mathcal{D}_{\text{indom}}$ is the FEM-pre-validated design pool of Figure 1(b): the generator arms produce 4,608 geometries on the 18 VF-stratified targets of $\mathcal{D}_{\text{base}}$ (6 from the training split, 12 from the validation split), ranked by the NOT proxy and validated by Abaqus FEM (protocol in Appendix A.3), each carrying an FEM ground-truth SS label; 4,597 are metric-valid after geometric and metric screening. The pool is split 8:1:1 into $\mathcal{D}^{\text{train}}_{\text{indom}}$ (3,675 usable), $\mathcal{D}^{\text{val}}_{\text{indom}}$ (459), and $\mathcal{D}^{\text{test}}_{\text{indom}}$ (463); its training split serves as the adaptation pool (Section 4.3), and the pool shares no geometry with the SDF dataset or the independent test set.

Under the $sdf < 0$ volume-fraction convention, the three splits share a similar VF mean while differing in spread (Figure 2): training (0.423 ± 0.058) and validation (0.426 ± 0.058) nearly coincide, whereas the independent test set reaches 0.422 ± 0.070 with extended tails. The independent set therefore differs from $\mathcal{D}_{\text{base}}$ by a distributional rather than a mean shift, which is the regime the extrapolation evaluation targets.

Disposition and failure-cause analysis of the anomalous target. Of the 12 evaluation targets in $\mathcal{D}^{\text{test}}_{\text{base}}$, one (source geometry: row 662 of $\mathcal{D}^{\text{test}}_{\text{base}}$) has an Abaqus reference SS curve that a provenance audit classifies as a data-level outlier and is therefore excluded, leaving $\mathcal{D}_{\text{indep}}$ with 11 targets. The audit establishes three findings: (i) the curve ends in a four-point upturn of $+106\%$ (last point against the fourth-from-last), the largest terminal rise among all 1,000 curves (next largest 51%; only $16/1{,}000$ exceed 20%); (ii) its maximum stress of 3.49 lies in the global 0.4% percentile and is the softest among the 151 cases in the same VF interval $[0.34, 0.38]$ (interval median 16.4), with this VF-0.362 target being $2.5\times$ softer than a VF-0.306 target — an inversion of the VF–stiffness relation; (iii) a pipeline origin is ruled out, as the v2→merged diff leaves this row unchanged (the 24 modified rows are all failure reruns) and the SDF solid fraction 0.3619 matches the manifest within a single connected domain. The anomaly thus originates in the FEM curve extraction of that batch and is judged irreparable. **This target is not a case of model failure**: across 12 model arms and 6 extrapolation levels all 240 attempts fail, the best candidate reaching $E_{\text{target-FEM}} = 23\%$ (median 202%) against a 5% criterion, so no model could pass under this reference curve. Excluding it leaves every ranking conclusion intact: under both the 11-target and the 12-target accounting, B2 ranks first, the B3 family ranks last, and r_{30} is the worst level.

Symbols. $\mathcal{D}_{\text{base}}$: the base data domain, comprising the original 10,000 geometries ($\mathcal{D}^{\text{train}}_{\text{base}}$/$\mathcal{D}^{\text{val}}_{\text{base}}$, 8,000/2,000) and the independently generated 1,000-geometry test split $\mathcal{D}^{\text{test}}_{\text{base}}$; $\mathcal{D}_{\text{indom}}$: the in-domain dataset, built on the 18 VF-stratified targets of $\mathcal{D}_{\text{base}}$ (6 from the training split, 12 from the validation split), 4,597 usable geometries; $\mathcal{D}^{\text{train}}_{\text{indom}}$/$\mathcal{D}^{\text{val}}_{\text{indom}}$/$\mathcal{D}^{\text{test}}_{\text{indom}}$: its random 8:1:1 splits (3,675/459/463), the test split (463 cases) used for proxy evaluation (Section 3.3); $\mathcal{D}_{\text{indep}}$: the independent test set, built on 11 targets of $\mathcal{D}^{\text{test}}_{\text{base}}$.

- $\mathcal{D}_{\textbf{base}}$: base data domain; the original 10,000 geometries split group-disjointly into 8,000 training and 2,000 validation, with 18 targets (6 in the training split, 12 in the validation split); plus the independently generated 1,000-geometry test split
- $\mathcal{D}_{\textbf{indom}}$: in-domain dataset, 18 targets; 4,608 generated geometries screened to 4,597 usable; split 8:1:1 into $\mathcal{D}^{\text{train}}_{\text{indom}}$=3,675, $\mathcal{D}^{\text{val}}_{\text{indom}}$=459, $\mathcal{D}^{\text{test}}_{\text{indom}}$=463
- **Selection pool**: 3,675 samples ($\mathcal{D}^{\text{train}}_{\text{indom}}$, metric-valid)
- **Independent test set**: $\mathcal{D}^{\text{test}}_{\text{base}}$, 1,000 geometries; VF sweep on 12 targets in two VF-stratified panels, primary protocol covers 11 targets (2,948 design-denominator FEM cases) after a provenance audit excluded one anomalous panel-2 target (q10_25, $\mathcal{D}^{\text{test}}_{\text{base}}$ row 662)

A.2 Independent Test Set Construction

A.2.1 Geometry Generation Protocol

The original release of Liu et al. (2025) provides only training and validation geometries; to evaluate generalization beyond the generator- and proxy-training distribution, the independent test set is generated from the full design space without any model involvement. Each geometry is drawn by the same random-field construction that underlies the original dataset: a periodic Gaussian random field is sampled on a 120×120 structured grid that covers the unit cell with a 10% margin (variance 10, correlation length 0.15, 32 Fourier modes); the field is thresholded at a contour level; and the solid phase is extracted and its boundary assembled with the original pipeline, which we replicate faithfully: boundary edge construction $\rightarrow$ greedy edge matching $\rightarrow$ small-hole filtering $\rightarrow$ near-point merging $\rightarrow$ narrow-region checking.

A single contour level per field, as in the original release, accepts only 0.5% of fields (measured over a 10-geometry pilot); searching up to 20 contour levels per field, drawn uniformly between the field's mid-range and 40% of its range above it, raises this to 4.4%. The full run yielded $1{,}000$ valid geometries in 91.8 minutes at a field-rejection rate of 96.2%, with volume fractions spanning $[0.226, 0.613]$ (mean 0.422), matching the original training distribution (Table 3). Every geometry is then labeled with an Abaqus SS curve (Appendix A.3): 980 completed in the first batch, and a supplementary batch of 25 geometries supplied replacements for the 20 solver failures and the 4 lowest-stress buckling cases; 6 cases with one unmatched periodic pair were retained because their curves are valid. The resulting set comprises $1{,}000$ geometries with complete FEM labels.

A.2.2 Evaluation Targets

The 12 evaluation targets are cases of $\mathcal{D}^{\text{test}}_{\text{base}}$ selected by volume fraction: six VF-percentile strata (0–10, 10–25, 25–50, 50–75, 75–90, and 90–100) each contribute one target in two independent panels. A provenance audit excludes one panel-2 target, leaving the 11 targets of the primary protocol (Appendix A.1).

A.2.3 Distribution Comparison

Table 3: Geometry-distribution comparison between the independent test set ($\mathcal{D}^{\text{test}}_{\text{base}}$, 1K) and the original SDF dataset (10K). VF follows the $sdf < 0$ convention; SDF range reports the global min/max over each split in un-normalized (stored) field values.

Statistic	Independent (1K)	Original (10K)
VF range	[0.226, 0.613]	[0.262, 0.553]
SDF range	[−0.369, 0.719]	[−0.250, 0.380]

The independent set exhibits a wider VF range ($[0.226, 0.613]$ vs $[0.262, 0.553]$) and broader SDF range, ensuring coverage of geometries outside the model's training distribution.

A.3 FEM Abaqus Configuration

Plane-strain compression with periodic boundary conditions on all four edges. Material: an elasto-plastic model (linear elasticity with $E = 2.306 \times 10^3$ and $\nu = 0.35$ in normalized units, plus a metal-plasticity hardening table; full table in the released script) under large deformation (nlgeom=ON). Elements: plane-strain CPE4 (CPE3 in degenerate regions) on a free adaptive mesh of nominal size 0.04. Contact: penalty formulation with slip fraction 0.005. Analysis step: static with automatic incrementation (initial 0.01, min 10^{-5}, max 0.1). Each simulation produces a 51-point stress-strain curve.

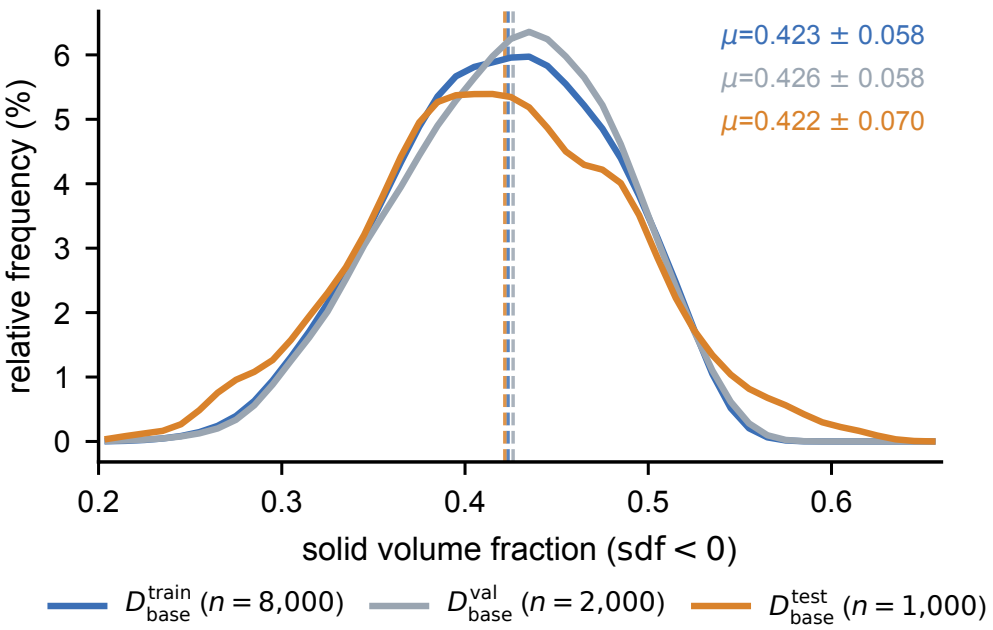


Figure 2: VF distribution of the three data splits under the $sdf < 0$ volume-fraction convention: $\mathcal{D}^{\text{train}}_{\text{base}}$ (8,000), $\mathcal{D}^{\text{val}}_{\text{base}}$ (2,000), and the independent test set (1,000). Y-axis: relative frequency (%).

A.4 Evaluation Protocol and Metrics

A.4.1 Unified Inference Protocol

All SS-VF generator arms and all targets share one inference protocol. The conditioning vector is formed by a target SS curve and a target VF, the latter set from the source geometry's $\text{VF}_{\text{source}}$ and a reduction level r as

$$\text{VF}_{\text{target}} = \text{VF}_{\text{source}} \times (1 - r), \quad r \in \{0,\ 0.10,\ 0.15,\ 0.20,\ 0.25,\ 0.30\}, \tag{14}$$

giving six levels. For each (arm, target, level) triple we generate 500 candidate SDFs, reject geometrically invalid ones with the periodicity and no-island gates, rank the survivors by proxy SS fidelity, and send the top-4 to Abaqus FEM for SS ground truth.

A.4.2 Metrics and Pass Criteria

- **Strict FEM pass:** A candidate passes if the Abaqus solver completes, the 51-point SS curve is valid, $E_{\text{target-FEM}} \leq 0.05$, and boundary validation passes (contour connectivity, periodic boundary condition node matching, bounding box within $[0, 1]^2$). The 5% threshold is adopted as the primary criterion throughout this work.
- **VF pass:** On FEM-passed candidates, the achieved VF extrapolation ratio must satisfy an asymmetric tolerance: geometries lighter than requested are always accepted, while geometries heavier than requested are accepted only within $|r_{\text{ach}} - r| \leq 0.02$ (Eq. (6)). This asymmetric threshold reflects the engineering preference for lighter structures and the empirical observation that VF deviations in FEM-passed candidates do not exceed 2% in the heavier direction.
- **pass@4 (top-4≥1):** For each target-level group, at least one of the top-4 ranked candidates passes the per-case criterion ($E_{\text{target-FEM}} \leq 5\%$ + asymmetric VF tolerance). This is the task-level success metric.

- $\boldsymbol{E}_{\text{NOT-FEM}}$: Relative L2 error between the NOT-predicted SS curve and the FEM ground truth, measuring the proxy-to-simulation gap. For samples whose FEM stress norm is exceptionally small ($\|\text{SS}_{\text{FEM}}\|_2 \approx 21$ vs. ≈ 231 for typical cases), the relative error can exceed 100% due to denominator amplification; these cases correspond to compliant, low-stiffness geometries with weak load paths.

B Model Details and Ablations

B.1 U-Net Backbone Configuration

Table 4: U-Net backbone configuration of the generator noise estimator (shared across the B1, B2, B3 conditioning families).

Hyperparameter	Value
Input shape	$1 \times 120 \times 120$ (SDF grid)
Total timesteps T	500
β-schedule	Linear ($\beta_1 = 2\times10^{-4}$, $\beta_T = 4\times10^{-2}$)
Base channels	16
Channel multipliers	$[1, 2, 4, 8]$
Residual blocks per level	1
Normalization	GroupNorm (groups = 8)
Self-attention levels	Levels 3, 4 (deepest two)
Time embedding	Sinusoidal $\rightarrow$ Linear

Training. All 12 generator arms are trained for 300 epochs with Adam (learning rate 10^{-3}), batch size 64. Classifier-free guidance (CFG) training sets the conditioning vector to empty with probability $p_{\text{uncond}} = 0.1$.

B.2 NOT Proxy Architecture Details

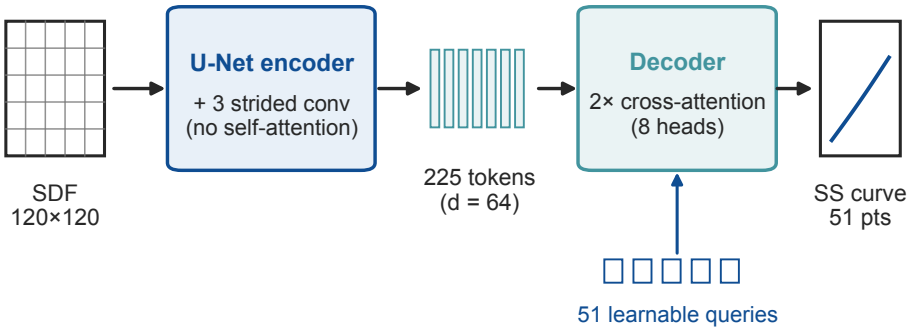


Figure 3: Neural Operator Transformer (NOT) proxy architecture.

Encoder. A U-Net backbone (Table 5) processes the 120×120 SDF into a feature map (Figure 3); it follows the generator layout but uses a base channel width of 8 instead of 16 and disables self-attention at every level (`has_attention=[F,F,F,F]`). Three successive Conv2d(stride $= 2$) + SiLU layers reduce the spatial resolution to 15×15 and project to an embedding dimension $d_{\text{emb}} = 64$. The feature map is flattened into a sequence of 225 tokens of dimension d_{emb}, serving as the context for cross-attention.

Decoder. A learnable query embedding of shape $(51, d_{\text{emb}})$ represents the 51 strain points. The query is projected through a linear layer and attends to the encoder tokens via $L_{\text{cross}} = 2$ cross-attention layers with $n_{\text{heads}} = 8$. The output is a 51-point SS curve. The total parameter count is approximately 923K.

Table 5: U-Net encoder backbone configuration of the NOT proxy. Rows mirror the generator configuration of Table 4.

Hyperparameter	Value
Input shape	$1 \times 120 \times 120$ (SDF grid)
Base channels	8
Channel multipliers	$[1, 2, 4, 8]$
Residual blocks per level	1
Normalization	GroupNorm (groups = 8)
Self-attention levels	None (disabled at all levels)
Time conditioning	None (deterministic regressor; no timestep embedding or β-schedule)

B.3 VF Loss Weight: Mechanism and Ablation

B.3.1 Implementation Details

The auxiliary VF loss is an MSE penalty between the soft VF of the one-step denoised prediction, $\hat{\text{VF}} = \mathcal{V}_\tau(\hat{x}_0)$, and the actual VF of the corresponding ground-truth SDF, $\text{VF}_{\text{actual}} = \mathcal{V}(x_0)$; both quantities are computed on the physical-unit SDF fields obtained by inverting the input normalization. The loss operates only on timesteps $t \leq t_{\max} = 100$ (late-stage, low-noise timesteps of the reverse process), with temperature $\tau = 0.005$. The total training objective combines the noise prediction loss L_ϵ (Eq. (9)) with the VF loss:

$$L = L_\epsilon + \lambda_{\text{vf}} L_{\text{VF}}. \tag{15}$$

$$L_{\text{VF}} = \text{MSE}(\hat{\text{VF}},\ \text{VF}_{\text{actual}}). \tag{16}$$

At $\lambda_{\text{vf}} = 0.01$, the VF loss contributes approximately 1% of the total loss, providing a weak gradient signal that enhances conditional awareness without precise VF control.

B.3.2 B1-VFLOSS Ablation Results

Table 6 reports the in-domain per-case pass rate for the VF loss weight comparison, over the same 18 VF-stratified targets and criterion as Table 8. The B1-VFLOSS-w05 variant is competitive at shallow reductions, reaching 50.0% at r_{15} against 45.8% for B1-VFLOSS-w01, yet degrades at depth: 34.7% at r_{20} and 18.1% at r_{30}, versus 37.5% and 30.6% for B1-VFLOSS-w01 and 37.5% and 33.3% for B1; over all levels it reaches 35.9%, below 40.7% for B1-VFLOSS-w01. The same ordering holds at the task level (pass@4 64.8% vs. 70.4%). The degradation is not domain-specific: on the independent set w05 reaches 18.2% at r_{30} (Table 12), matching its in-domain 18.1%.

Table 6: VF loss weight ablation. $\mathcal{D}_{\text{indom}}$: per-case pass rate (%), 72 cases per cell (18 targets $\times$ top-4); B1 is the $\lambda_{\text{vf}} = 0$ reference.

Model	r_0	r_{10}	r_{15}	r_{20}	r_{25}	r_{30}	ALL
B1	41.7	51.4	47.2	**37.5**	**27.8**	**33.3**	39.8
B1-VFLOSS-w01 ($\lambda = 0.01$)	**50.0**	**54.2**	45.8	**37.5**	26.4	30.6	**40.7**
B1-VFLOSS-w05 ($\lambda = 0.05$)	43.1	47.2	**50.0**	34.7	22.2	18.1	35.9

B.4 B1 Ablation: Cross-Dataset Comparison

B1, B1-D (conditioning dropout), and B1-PD (physics dropout) share the B1 plain-concatenation backbone and differ only in their dropout mechanism. We compare them across two evaluation regimes: in-distribution ($\mathcal{D}_{\text{indom}}$) and the independent test set ($\mathcal{D}_{\text{indep}}$).

B1-PD leads the B1 family on the 11-target independent test set (35.6% vs 34.1% per-case, 71.2% vs 66.7% top-4$\geq$1), whereas the in-distribution $\mathcal{D}_{\text{indom}}$ ranking is comparatively flat and in the opposite direction (B1 39.8%, B1-D 38.9%, B1-PD 33.8%). The per-case values are the aggregated columns of Tables 1 ($\mathcal{D}_{\text{indep}}$) and 8 ($\mathcal{D}_{\text{indom}}$); the top-4 values are those of Tables 13 and 9.

B.5 Inference Speed Benchmark

Controlled single-GPU benchmark (RTX 4090, 500 denoising timesteps):

Table 7: Parameters and inference speed by conditioning family.

Family	Params	Speed (cand/s)	Time for 500 candidates
B1 (concat)	4.190 M	2.45	204 s
B2 (xattn+FiLM)	2.326 M	3.14	159 s
B3 (concat+FiLM)	2.065 M	3.45	145 s

Speed differences arise from the conditioning mechanism, not backbone variants (D/PD/VFLOSS share speed within family). Parameter counts are read directly from the released checkpoints, and the conditioning modules differ in capacity across families (B1 2.078 M / B2 0.222 M / B3 0.033 M). We report these differences in condition-module capacity rather than claiming parameter matching; all arms share an identical training budget (300 epochs, same batch size and data). The smaller B2/B3 condition modules nonetheless attain higher pass@4 than B1 (Section 4.2), indicating that the gains stem from the conditioning mechanism rather than capacity stacking.

C Additional Evaluation Matrices

C.1 In-Domain 18-Target Matrix

This subsection reports the in-domain 18-target matrix at the per-case (Table 8) and pass@4 (Table 9) levels. The SS-only control B0 is defined only at r_0, and — marks cells where an arm is not defined.

Table 8: $\mathcal{D}_{\text{indom}}$: per-case pass rate (%), 72 cases per cell (18 targets × top-4).

Family	Model	r_0	r_{10}	r_{15}	r_{20}	r_{25}	r_{30}	ALL
SS-only	B0	**56.9**	—	—	—	—	—	56.9
B1 (concat)	B1	41.7	51.4	47.2	37.5	27.8	33.3	39.8
	B1-D	45.8	45.8	50.0	31.9	30.6	29.2	38.9
	B1-PD	37.5	41.7	41.7	36.1	23.6	22.2	33.8
	B1-VFLOSS-w01	50.0	54.2	45.8	37.5	26.4	30.6	40.7
B2 (xattn+FiLM)	**B2**	37.5	**59.7**	38.9	38.9	**37.5**	**38.9**	**41.9**
	B2-D	40.3	43.1	37.5	41.7	26.4	27.8	36.1
	B2-PD	45.8	59.7	48.6	27.8	38.9	25.0	41.0
B3 (concat+FiLM)	B3	36.1	23.6	26.4	16.7	13.9	5.6	20.4
	B3-D	27.8	29.2	20.8	11.1	18.1	13.9	20.1
	B3-PD	27.8	15.3	22.2	11.1	11.1	5.6	15.5

Table 9: $\mathcal{D}_{\text{indom}}$: pass@4 (top-4≥1) pass rate (%), requiring the per-case criterion (Appendix A.4), over 108 target-level groups (18 targets × 6 levels).

Family	Model	r_0	r_{10}	r_{15}	r_{20}	r_{25}	r_{30}	ALL
SS-only	B0	**88.9**	—	—	—	—	—	88.9
B1 (concat)	B1	77.8	**100.0**	77.8	61.1	66.7	55.6	73.1
	B1-D	83.3	77.8	77.8	61.1	50.0	44.4	65.7
	B1-PD	72.2	77.8	72.2	55.6	55.6	50.0	63.9
	B1-VFLOSS-w01	88.9	77.8	77.8	66.7	50.0	61.1	70.4
B2 (xattn+FiLM)	**B2**	77.8	88.9	66.7	72.2	66.7	66.7	**73.1**
	B2-D	66.7	72.2	72.2	66.7	55.6	50.0	63.9
	B2-PD	83.3	88.9	77.8	55.6	61.1	55.6	70.4
B3 (concat+FiLM)	B3	72.2	55.6	50.0	38.9	38.9	16.7	45.4
	B3-D	55.6	50.0	44.4	33.3	44.4	27.8	42.6
	B3-PD	61.1	50.0	38.9	27.8	38.9	16.7	38.9

C.2 Evaluation Optimism across Test Sets

Table 10: Diagnosing evaluation optimism: surrogate proxy accuracy on model-generated vs. independent OOD geometries.

Evaluation set	n	$E_{\text{NOT-FEM}}$ mean	Source
$\mathcal{D}^{\text{test}}_{\text{indom}}$ split (model-generated)	463	0.062	Generator output
$\mathcal{D}_{\text{indep}}$ sweep (model-generated)	2,924	0.077	Generator output
1K independent test set (full design space)	**1,000**	**0.174**	**OOD original**

C.3 Aggregate Metrics by VF Extrapolation Level

Table 11: Aggregate metrics by extrapolation level on $\mathcal{D}_{\text{indep}}$. r_0 pools all 12 arms (B0 participates only at r_0); r_{10}–r_{30} pool the 11 VF-aware arms.

Red.	n	$E_{\text{t-NOT}}$	$E_{\text{t-FEM}}$	Pass Rate (%)
r_0	528	0.029	0.087	37.5%
r_{10}	484	0.040	0.083	35.1%
r_{15}	484	0.050	0.096	36.4%
r_{20}	484	0.064	0.102	35.1%
r_{25}	484	0.090	0.119	27.7%
r_{30}	484	0.109	0.139	20.7%
ALL	2,948	—	0.104	32.2%

C.4 Independent Test Set: Full Model Matrix

Trajectory morphology and extrapolation mechanism.

For the B1 and B2 families, per-case pass rates peak at shallow extrapolation levels, typically r_{10} in-domain and r_{10}–r_{20} on the independent test set, and broadly decline toward r_{30}, with per-arm fluctuation at r_{25}–r_{30}. This morphology is consistent across domains, indicating that it reflects the conditioning mechanism rather than the evaluation distribution. It suggests that the VF at r_0 may not be the comfortable operating point of the given SS target; at moderate extrapolation levels the generator produces more diverse geometries, and a light downward VF shift grants additional freedom to rearrange material within the existing structural prototype. We state this as a hypothesis rather than a general law. The B3 family is excluded from the trajectory analysis because its overall performance is far too weak for trajectory-level conclusions to be meaningful; its r_0-peaked decline reflects overall failure rather than a distinct extrapolation mechanism, and its role as the conditioning-architecture negative control is reported in Table 1. Figure 4 and Figure 5 show the candidate-level and task-level trajectories for the four leading B1/B2 arms; shaded bands mark Wilson 95% confidence intervals for the two leading arms (B2 and B2-PD), and the light gray region marks the decline regime beyond r_{20}.

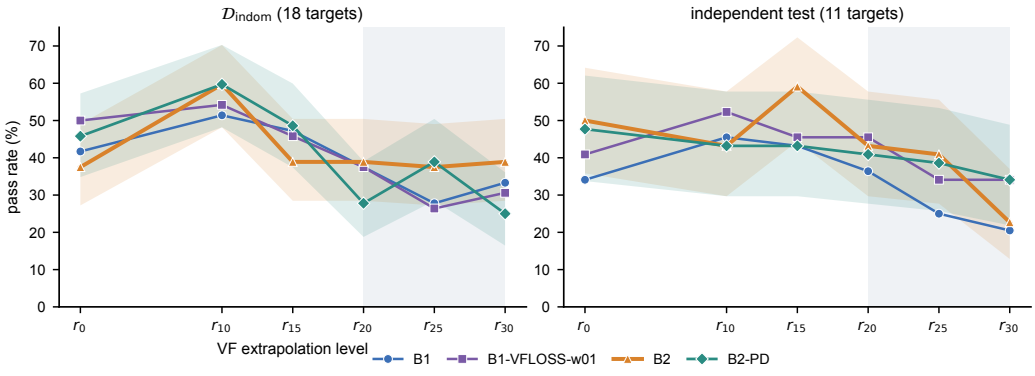


Figure 4: Candidate-level per-case pass rates under VF extrapolation. **Left:** in-domain $\mathcal{D}_{\text{indom}}$ (18 targets; $n = 72$ per cell). **Right:** independent-test $\mathcal{D}_{\text{indep}}$ (11 targets; $n = 44$ per cell). Lines: B1, B1-VFLOSS-w01, B2, B2-PD. The task-level counterpart is shown in Figure 5.

Qualitative exemplar under deep extrapolation. The qualitative panels use a median-severity exemplar target, selected deterministically among the targets where B2 passes at r_0 but fails at r_{30}; at r_{30} even the best B2 candidate misses the threshold ($E_{\text{target-FEM}} = 7.3\%$), and the solid phase

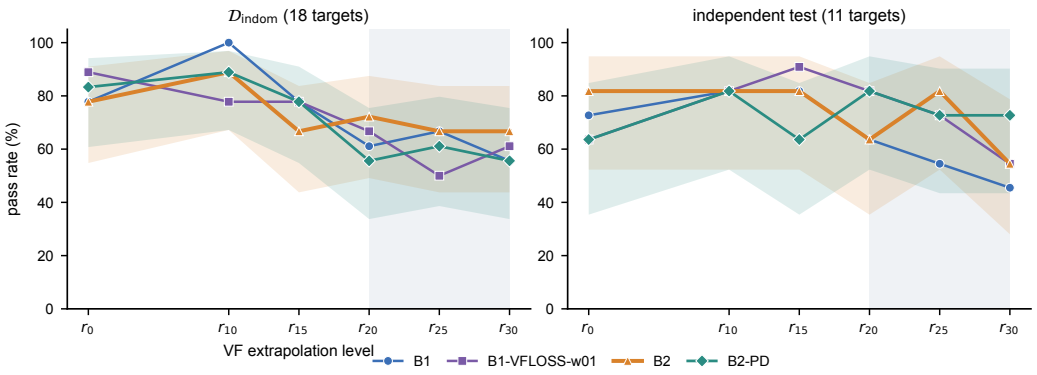


Figure 5: Task-level pass@4 rates under VF extrapolation. **Left:** in-domain $\mathcal{D}_{\text{indom}}$ ($n = 18$ target-level groups). **Right:** independent-test $\mathcal{D}_{\text{indep}}$ ($n = 11$). Lines and shading as in Figure 4.

becomes progressively fragmented along the decline regime of Figure 4, with B3 degrading earlier and B1 remaining mixed.

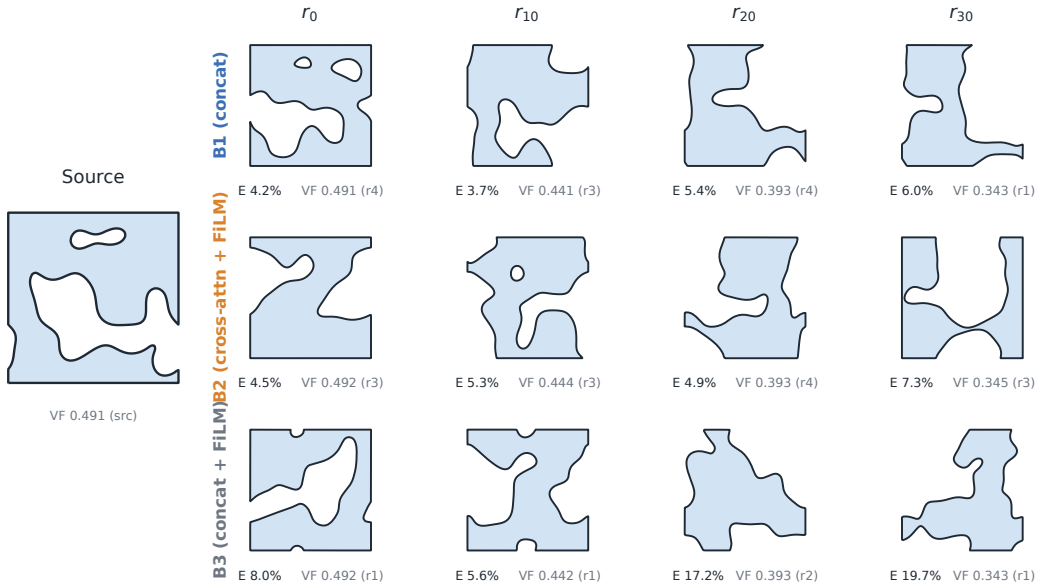


Figure 6: Qualitative SDF generation under VF extrapolation ($\mathcal{D}_{\text{indep}}$; zero level sets, solid phase in light blue). Rows: architectures B1/B2/B3; columns: VF levels. Each panel shows the top-4 candidate with the best FEM E_{target} (rank annotated) and its $E_{\text{target-FEM}}$ and achieved VF; the exemplar target is a median-severity pick, chosen deterministically (Section C.4).

Table 12: $\mathcal{D}_{\text{indep}}$: per-case pass rate (%) for the auxiliary-VF-loss variants, inverse error $E_{\text{target-FEM}} \leq 5\%$ + asymmetric VF tolerance $\epsilon = 0.02$ (criterion identical to Appendix A.4). B1 is repeated as the family reference. w01 denotes $\lambda_{\text{vf}} = 0.01$; w05 denotes $\lambda_{\text{vf}} = 0.05$.

Model	r_0	r_{10}	r_{15}	r_{20}	r_{25}	r_{30}	ALL
B1 (reference)	34.1	45.5	43.2	36.4	25.0	20.5	34.1
B1-VFLOSS-w01	40.9	52.3	45.5	45.5	34.1	34.1	42.0
B1-VFLOSS-w05	40.9	52.3	47.7	29.5	29.5	18.2	36.4

Table 13: $\mathcal{D}_{\text{indep}}$: pass@4 (top-4≥1) pass rate (%), requiring the per-case criterion (Appendix A.4), over 66 target-level groups (11 targets × 6 levels). B0 is an SS-only control defined only at r_0; its r_0 column is directly comparable with the other arms, while its ALL column aggregates the 11 r_0 target-level groups only.

Family	Model	r_0	r_{10}	r_{15}	r_{20}	r_{25}	r_{30}	ALL
SS-only	B0	63.6	—	—	—	—	—	63.6
B1 (concat)	B1	72.7	81.8	81.8	63.6	54.5	45.5	66.7
	B1-D	54.5	72.7	63.6	81.8	54.5	45.5	62.1
	B1-PD	72.7	72.7	81.8	63.6	72.7	63.6	71.2
	B1-VFLOSS-w01	63.6	81.8	90.9	81.8	72.7	54.5	74.2
	B1-VFLOSS-w05	63.6	90.9	81.8	45.5	63.6	36.4	63.6
B2 (xattn+FiLM)	**B2**	**81.8**	81.8	**81.8**	63.6	**81.8**	54.5	**74.2**
	B2-D	63.6	72.7	72.7	81.8	63.6	45.5	66.7
	B2-PD	72.7	81.8	63.6	81.8	72.7	**72.7**	**74.2**
B3 (concat+FiLM)	B3	63.6	18.2	27.3	45.5	27.3	18.2	33.3
	B3-D	45.5	54.5	36.4	36.4	27.3	27.3	37.9
	B3-PD	63.6	72.7	36.4	45.5	18.2	18.2	42.4

D NOT PROXY ADAPTATION: MECHANISMS AND ABLATIONS

This appendix consolidates the adaptation-mechanism studies referenced in Section 4.3: the adaptation trainers, residual-guided selection ablations, the per-sample distribution of selection improvement, the selection $\times$ accumulation decomposition, and the pool-size efficiency of the adapted ranker. Unless stated otherwise, results are 3-seed means; the independent test set contains 1,000 geometries and the $\mathcal{D}^{\text{test}}_{\text{indom}}$ split 463 cases.

D.1 ADAPTATION TRAINERS

We compare the two adaptation trainers under fixed fine-tuning protocols; the in-loop trainer swap (S1 adapter vs. S2 fine-tuning with anchoring) is reported in Table 2, and the selection results there address sample selection under a round budget.

Trainer definitions.

- **S1 (frozen-backbone adapter).** The NOT backbone is frozen; a lightweight residual adapter ($h = 128$ hidden units, 37K parameters) is trained on FEM-labeled data, inserted after the decoder output.
- **S2 (fine-tuning).** The NOT backbone is unfrozen and fine-tuned on FEM-labeled data with a pure MSE objective.

D.2 RESIDUAL-GUIDED SELECTION

Algorithm 1 Residual-guided selection (SFV) loop.

Require: teacher proxy $\mathcal{N}_{\theta_T}$; $\mathcal{D}^{\text{train}}_{\text{base}}$; pre-validated pool $\mathcal{D}_{\text{pool}} = \{(x_i, y_i^{\text{FEM}})\}$ (i.e., $\mathcal{D}^{\text{train}}_{\text{indom}}$, 3,675 cases); batch n_{sel}; anchor weight λ; budget R

1: $\theta^{(0)} \leftarrow \text{FineTune}(\theta_T)$ on $\mathcal{D}^{\text{train}}_{\text{base}}$ with the ℓ_2 anchor (R_0; teacher-anchored)
2: **for** round $r = 1, \ldots, R$ **do**
3: **Select:** score remaining pool with the current student, $\epsilon_i = \|\mathcal{N}_{\theta^{(r-1)}}(x_i) - y_i^{\text{FEM}}\|_2 / \|y_i^{\text{FEM}}\|_2$; select top-$n_{\text{sel}}$ by ϵ (residual mode) or uniformly at random (random mode), without replacement; if fewer than n_{sel} remain, all are selected (the final round may be partial)
4: **Fit:** warm-start from $\theta^{(r-1)}$; fine-tune on $\mathcal{D}^{\text{train}}_{\text{base}} \cup \mathcal{D}^{(r)}_{\text{sel}}$, where $\mathcal{D}^{(r)}_{\text{sel}}$ is all pool samples selected so far (cumulative) or only the current round's batch (non-cumulative), with the anchored objective in Eq. (13)
5: **Validate:** within each round, select the round's checkpoint by validation loss (early stopping, patience 20)
6: **end for**
7: **return** the round-level val-best checkpoint (best validation loss over rounds $1, \ldots, R$); evaluate once on $\mathcal{D}^{\text{test}}_{\text{base}}$ and $\mathcal{D}^{\text{test}}_{\text{indom}}$

All experiments use a budget of at most 10 rounds (terminating early if the pool is exhausted), the val-best reporting protocol, and 3 seeds. $\mathcal{D}^{\text{test}}_{\text{base}}$ (1,000 independent FEM cases) serves as the primary evaluation; the $\mathcal{D}^{\text{test}}_{\text{indom}}$ split (463 cases) as auxiliary. Figure 7 shows the dual-trajectory view ($\mathcal{D}^{\text{test}}_{\text{base}}$ and $\mathcal{D}^{\text{test}}_{\text{indom}}$) for the three selection arms: both evaluation sets improve simultaneously with no trade-off between them, the cumulative arm decreasing monotonically on $\mathcal{D}^{\text{test}}_{\text{indom}}$ but saturating on $\mathcal{D}^{\text{test}}_{\text{base}}$ after about three rounds. Algorithm 1 summarizes the full loop.

D.2.1 CUMULATIVE MODE: FULL MULTI-SEED RESULTS

Cumulative mode is λ-insensitive: all configurations achieve $+6.4\%$ to $+6.5\%$ on $\mathcal{D}^{\text{test}}_{\text{base}}$ (differences within std). $\mathcal{D}_{\text{indom}}$ growth across rounds provides natural regularization, rendering the explicit L2 anchor less critical. The recommended configuration is Cum 400/round $\lambda = 2.0$, which minimizes the validation loss across the grid (mean 0.00341, seed spread 6×10^{-6}) and is therefore selected on validation; its test-set results are reported once above.

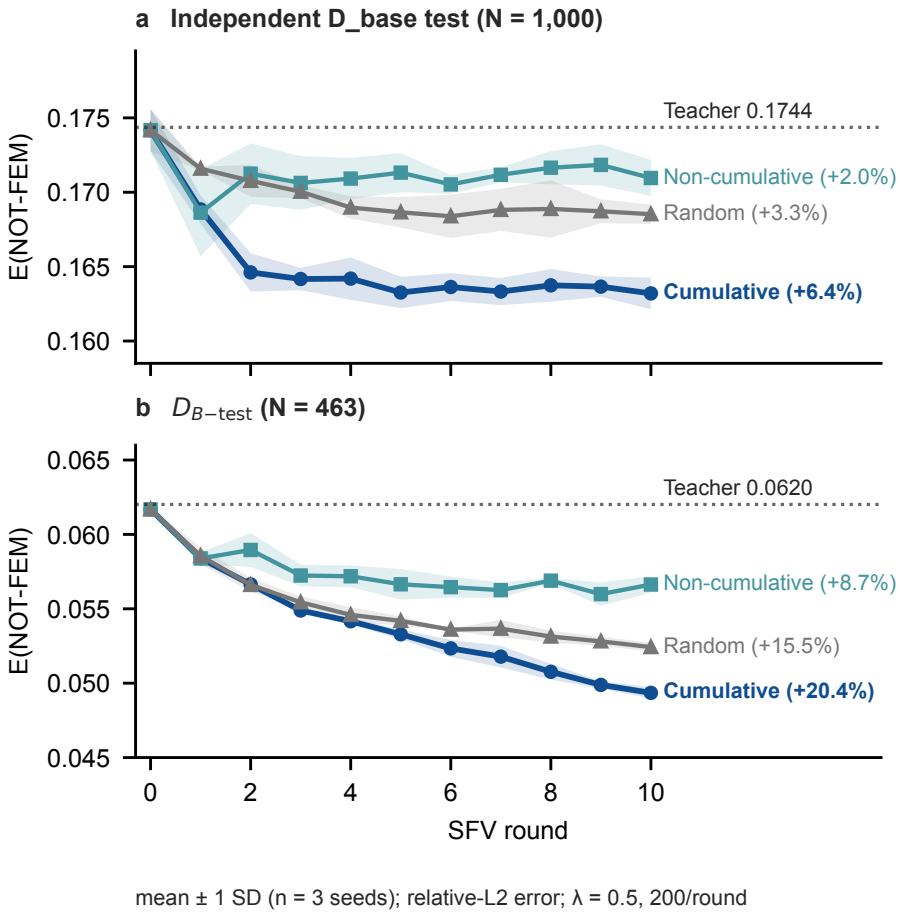


Figure 7: Selection trajectory over the 10-round budget (3 seeds, mean $\pm$ 1 SD; $\lambda = 0.5$, 200 samples/round). (a) $\mathcal{D}^{\text{test}}_{\text{base}}$ independent test set (1,000 cases); (b) $\mathcal{D}^{\text{test}}_{\text{indom}}$ split (463 cases). Lines: cumulative residual, non-cumulative, and random selection arms; dotted: frozen teacher baseline. Trajectory endpoints are the round-10 values, tabulated multi-seed results use val-best checkpoints (Section D.2).

Table 14: Cumulative mode multi-seed results (3 seeds, fixed 200 labels/round, 10 rounds, val-best checkpoints). $\mathcal{D}^{\text{test}}_{\text{base}}$ test = 1,000 independent FEM cases. The 400/round endpoints ($\lambda = 0.5$, 2.0) are reported in the main text and Table 2.

Configuration	Seeds	$\mathcal{D}^{\text{test}}_{\text{base}}$ mean$\pm$std	Impr.	$\mathcal{D}^{\text{test}}_{\text{indom}}$ mean$\pm$std	Impr.
Cum 200 λ=0.5	3	0.163 ± 0.0011	$+6.4\%$	0.049 ± 0.0004	$+20.4\%$
Cum 200 λ=1.0	3	0.163 ± 0.0004	$+6.5\%$	0.049 ± 0.0008	$+20.3\%$
Cum 200 λ=2.0	3	0.163 ± 0.0009	$+6.5\%$	0.050 ± 0.0003	$+20.0\%$

D.2.2 Selection Rule (Non-Cumulative, Multi-Seed)

In non-cumulative mode, random selection at $\lambda = 0.5$ outperforms residual-based selection ($+3.3\%$ vs $+2.0\%$ on $\mathcal{D}^{\text{test}}_{\text{base}}$; $+15.5\%$ vs $+8.7\%$ on $\mathcal{D}^{\text{test}}_{\text{indom}}$). The margin between the two rules is regularization-dependent: the sweep in Table 14 shows the gap narrowing as the L2 anchor strengthens. This indicates that the diversity advantage of random selection diminishes when stronger regularization counters the distributional shift from greedy residual selection. In cumulative mode this effect is ameliorated by the growing pool.

Table 15: Selection rule comparison (non-cumulative, MSE+L2, fixed $\lambda = 0.5$, 200/round, 10 rounds, 3 seeds, val-best checkpoints).

Rule	$\mathcal{D}^{\text{test}}_{\text{base}}$ mean±std	Impr.	$\mathcal{D}^{\text{test}}_{\text{indom}}$ mean±std	Impr.
Residual top-200	0.171 ± 0.0009	$+2.0\%$	0.056 ± 0.0007	$+9.2\%$
Random 200	0.169 ± 0.0007	$+3.3\%$	0.052 ± 0.0003	$+15.5\%$

Downstream screening utility. Finally, we verify the practical downstream effect on the generation pipeline. Table 16 re-ranks the four candidates of each of the 736 cells (11 targets) with the adapted proxy versus the frozen SS-only NOT; the cumulative-residual trainer raises the top-1 FEM-valid rate from 34.4% to 39.6% using no additional FEM simulations; candidates carry precomputed labels and re-ranking uses no labels. The deployed shortlists were themselves ranked by the same frozen SS-only NOT, whose stored order and recomputed re-ranking differ by one cell (34.5% vs 34.4% of 736; McNemar $p = 1.0$), so Table 16 reports a single baseline row.

Table 16: Downstream screening utility of proxy adaptation on $\mathcal{D}_{\text{indep}}$ (736 cells). Column *hit@1@5%*: the fraction of cells whose top-ranked candidate is FEM-valid at threshold $\tau = 5\%$, i.e. inverse error $E_{\text{target-FEM}} \leq \tau$ plus the asymmetric VF tolerance $\epsilon = 0.02$ (criterion identical to Appendix A.4). All rows re-rank the four candidates of each cell by the proxy score. The matched-configuration rows differ *only* in the Select-step selection rule (same pool, 200 labels/round, 10 rounds, 3 seeds, val-best checkpoints). Selection-rule comparison in Table 15; the S1 mechanism is analysed below.

Ranking / selection rule	Configuration	hit@1@5%
Ranking proxy		
frozen SS-only NOT	—	34.4
S1 adapter (frozen backbone)	one-shot	34.7
Selection rule		
random selection	200/round, $\lambda = 0.5$	36.7
non-cumulative residual	200/round, $\lambda = 0.5$	36.2
cumulative residual	200/round, $\lambda = 0.5$	39.3
cumulative residual (recommended)	400/round, $\lambda = 2.0$	39.6

To isolate *why* the closed loop helps, we re-rank the same 736 cells with the **S1 frozen-backbone adapter** (the trainer-ablation arm of Table 2), which reduces the out-of-domain proxy error *more* aggressively than the closed loop ($+14.3\%$ vs $+6.5\%$). Its top-1 FEM-valid rate is nevertheless indistinguishable from the frozen SS-only NOT (34.7% vs 34.4%; $+0.3$ pp; target-clustered bootstrap over the 11 targets, $p = 0.79$), whereas the closed loop reaches 39.6% ($+5.2$ pp, $p = 0.0016$) and is significantly better than S1 (-4.9 pp, $p = 0.0008$). The difference lies in *within-cell* ranking: among the four candidates of a cell, the Spearman correlation between the proxy score and the true $E_{\text{target-FEM}}$ is 0.082 (frozen SS-only NOT), 0.112 (S1) and 0.237 (closed loop). Lower absolute proxy error therefore does not by itself translate into screening yield — S1 barely reorders within a cell, while the FEM-anchored loop improves precisely the ordering that screening consumes, which is why adaptation is validated end-to-end on the generation pipeline rather than by proxy error alone.

D.3 Pool-Size Efficiency of the Adapted Ranker

Does proxy accuracy improve screening yield? Ranking quality determines how much FEM budget a fixed candidate pool requires. Holding the candidate pool fixed and varying *only* the ranking proxy, we subsample N' candidates per group from the $\mathcal{D}^{\text{test}}_{\text{indom}}$ split of the adaptation pool ($\mathcal{D}^{\text{test}}_{\text{indom}}$, 463 cases; 18 targets, 18–35 candidates each, all FEM-labelled), rank them by predicted error, and FEM-validate only the top-ranked candidate (k=1). The unadapted teacher peaks at 44.5% (at N'=4) and stays below 45% at every pool size tested ($N' \leq 18$). The adapted proxy reaches 48.4% at N'=3, already above the teacher's best attainable yield at any pool size, and 51.2% at N'=4; at the matched pool size N'=4, adaptation therefore raises the top-1 yield from 44.5% to 51.2%

(paired 95% CI over the 18 target groups: $+0.4$ to $+12.9$ percentage points). The advantage grows with pool depth, from $1.05\times$ at $N'=2$ to $1.53\times$ at $N'=18$. Because the candidate pools and the FEM labels are identical across rankers, this comparison isolates ranking accuracy as the cause; it is made at a matched pool size rather than read as a trend, since a single-validation top-1 yield is a precision statistic and is not monotone in N'.

Out-of-domain anchor. The same comparison can be run out of domain. On the 11 novel targets of Table 16 (a fixed pool of four candidates per cell, with zero overlap with the adaptation pool), adaptation raises the top-1 FEM-valid rate from 34.4% to 39.6% (paired 95% CI above zero). The relative gain at the shared pool size of four is nearly identical in and out of domain ($+15.1\%$ vs. $+15.2\%$): the adapted proxy loses little when transferred to unseen targets, indicating that the ranking advantage is not merely in-domain memorisation. Adaptation used only 400 FEM labels per round over at most 10 rounds.

D.4 Per-Sample Distribution of Selection Improvement

Per-sample change analysis ($R_0 \rightarrow R_{10}$, Cum $\lambda = 0.5$; per-geometry mean over 3 seeds, $N = 1{,}000$) reveals that residual-guided selection improvement concentrates on hard samples while degradation on easy samples is negligible. Figure 8 visualizes the transition between baseline-error bands; the band-change statistics are given below.

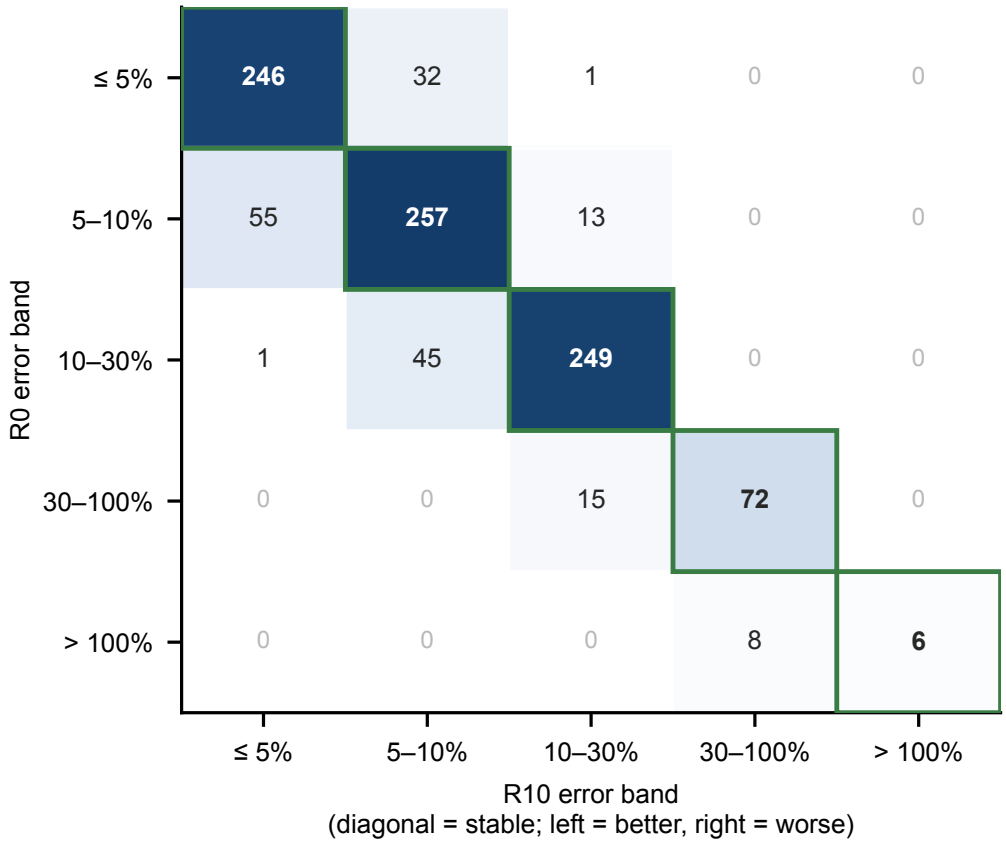


Figure 8: Error-band transition of per-sample proxy error under residual-guided selection ($R_0 \rightarrow R_{10}$, Cum $\lambda = 0.5$; per-geometry mean over 3 seeds, $N = 1{,}000$). Rows give the R_0 baseline-error band and columns the R_{10} band of the same geometry, with cell values as sample counts; the green diagonal marks stable samples that do not change band, cells left of the diagonal move to a better band and cells right of it to a worse band. The band boundaries (5%, 10%, 30%, 100%) are the only thresholds used.

No hard band degrades: none of the 396 geometries with $R_0 > 10\%$ moves to a worse band, and 46 of them (11.6%) cross into $\leq 10\%$. Degradation is confined to the two easy bands and totals 46 geometries (4.6%), a set disjoint from the 46 improving hard geometries: 33 from $\leq 5\%$ and 13 from 5–10% move to a worse band, yet the $\leq 5\%$ bin mean changes only from 3.13% to 3.35% ($\Delta = +0.0023$). The overall +6.3% mean improvement is driven by hard-sample correction and a reduction of extreme outliers ($> 100\%$ error falls from 14 to 6 geometries). This finding supports the residual-guided design logic: residual-based sampling prioritizes labeling the samples where the proxy is least accurate, yielding disproportionate gains on the hardest cases. Samples with $E_{\text{NOT-FEM}} > 100\%$ stem from the relative-error formula: the FEM stress norm of hard samples can be as low as 21 (vs. 231 for easy samples), so a vanishing denominator amplifies small absolute prediction errors into large relative ones. The most extreme case reaches 32×; this is a mathematical amplification effect on low-stress samples, not a complete prediction failure.

D.5 Selection × Accumulation Decomposition

Table 17: Selection × accumulation 2 × 2 in the lightweight adapter regime ($\mathcal{D}_{\text{indom}}$-only samples, no $\mathcal{D}_{\text{base}}^{\text{train}}$ mixing; round selected by $\mathcal{D}_{\text{indom}}$ validation; 3 seeds, mean ± SD, independent test set). Cumulative accumulation is the dominant factor; residual selection adds little on its own.

	Cumulative	Non-cumulative
Residual	$\mathbf{0.151 \pm 0.0013}$	0.175 ± 0.0029
Random	0.153 ± 0.0019	0.157 ± 0.0049

Under the $\mathcal{D}_{\text{indom}}$-validation round-selection convention used here, the main-protocol combined-val convention yields 0.167 ± 0.0071 for the cumulative-residual cell; the two conventions must be read separately. A frozen-backbone adapter trained once on the full pool reaches 0.149 ± 0.0004 ($\mathcal{D}_{\text{indom}}$ 0.057 ± 0.0002), essentially matching the 10-round loop endpoint (0.151): online selection brings no additional gain in the adapter regime.