FABBENCH: A FABRICATION-INFORMED BENCHMARK FOR SEM-DOMAIN PROCESS CONDITIONED LITHOGRAPHY MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurately predicting semiconductor lithography patterns and defect states under continuous process parameters is vital for advanced chip design. However, existing benchmarks either focus purely on physics simulation or isolate defect analysis from SEM domain observations. The scarcity of process-annotated wafer data further prevents joint evaluation of geometric evolution and defect formation under continuous process variations, leaving their coupled behavior largely unexamined. Moreover, conventional image-level metrics are poorly aligned with fabrication-side assessment, often obscuring localized metrological errors and failures in defect prediction. To address these limitations, we introduce FabBench, the first comprehensive benchmark for evaluating process-conditioned geometric evolution and defect formation in the SEM domain. Built on 547K physics-grounded, real-SEM-informed triplets with structured in-distribution and out-of-distribution splits, FabBench establishes a hierarchical, metrology-aware evaluation protocol spanning visual fidelity, localized dimensional accuracy, and topological defect consistency. Extensive evaluations show that predictions with high image-level similarity can still exhibit substantial errors under fabrication-aligned metrology and defect metrics, revealing a critical gap between visual fidelity and fabrication-relevant accuracy. FabBench thus provides a unified testbed for developing models that accurately capture both fine-grained geometric responses and topological defect formation across continuous process variations.

1 INTRODUCTION

As technology scaling advances, semiconductor design-technology co-optimization (DTCO) increasingly extends to design-to-manufacturing verification Zhang et al. (2023b). Accurately predicting lithography SEM images and defect formation under varying process conditions is now critical in computational lithography He et al. (2026); Wang et al. (2022). Consequently, establishing controlled and reproducible benchmarks is crucial to systematically evaluate whether computational models faithfully capture process-induced geometric evolutions and defect behaviors.

To address this challenge, prior research has largely diverged into two isolated paradigms. The first centers on physics-based lithography simulation Banerjee et al. (2013); Jin et al. (2025); Wang et al. (2025); He et al. (2025); Zheng et al. (2023), where benchmarks typically generate idealized binary resist contours under controlled process conditions. This paradigm provides a scalable and mathematically explicit mapping from continuous process parameters to physical responses. However, it remains confined to simulated representations and cannot fully capture the imaging artifacts, noise, and complex surface morphology observed in real-world scanning electron microscopy (SEM).

Conversely, SEM-domain defect analysis directly studies topological anomalies from physical wafer images. Existing datasets, however, predominantly formulate defect analysis as an isolated recognition or segmentation task, without paired process parameters that enable conditional forward modeling of how patterns evolve as fabrication conditions vary. This limitation is difficult to overcome because authentic wafer SEM data are scarce, inherently long-tailed, and rarely accompanied by dense process annotations. As a result, existing benchmarks face a fundamental trade-off: physics-driven simulation offers scale, continuous control, and counterfactual process variation but suffers

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.

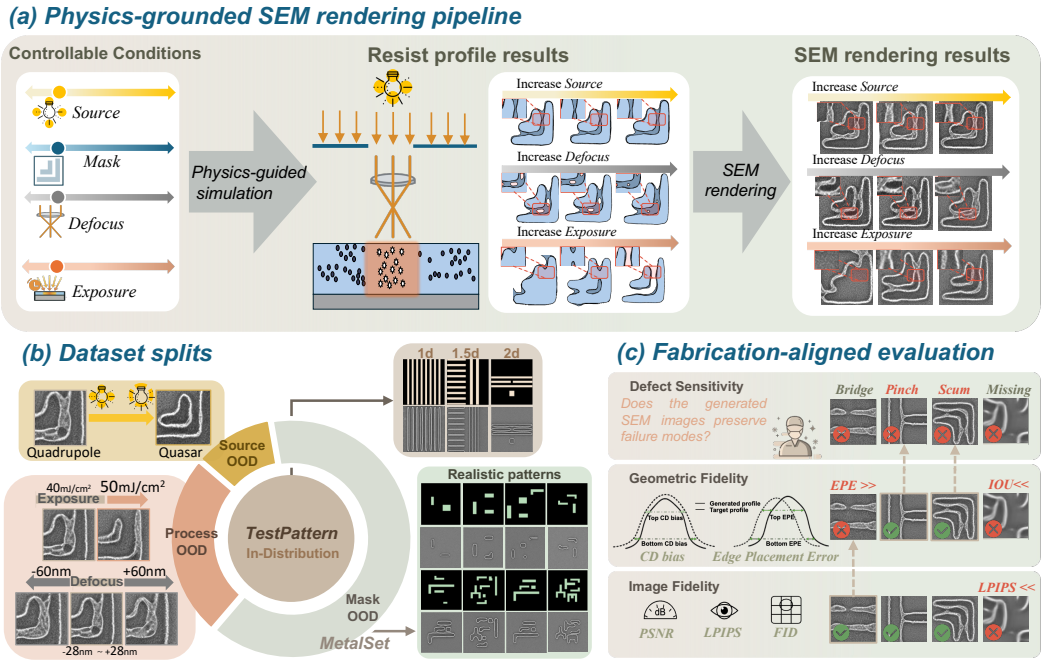


Figure 1: FabBench couples physics-grounded lithography simulation with SEM rendering to construct process-conditioned mask–process–SEM triplets under controllable source, defocus, and exposure conditions. Structured ID, Mask-OOD, Process-OOD, and Source-OOD splits evaluate generalization across mask topology and process variations, while fabrication-aligned metrics jointly assess image fidelity, local geometric accuracy, and topological defect formation.

from a substantial simulation-to-reality gap, whereas real-wafer datasets provide authentic SEM observations but lack the scale, process control, and paired annotations required to systematically evaluate process-conditioned geometric evolution and defect formation. Beyond this data bottleneck, existing evaluation protocols are also poorly aligned with fabrication-side assessment. Lithography prediction models are commonly evaluated using contour-overlap metrics such as IoU or generic image-similarity measures such as SSIM inherited from conventional computer vision benchmarks. While these metrics capture global pattern agreement, they do not directly assess whether a model reproduces fine-grained dimensional changes or correctly predicts process-induced topological defects. Consequently, predictions with high contour overlap or visual similarity may still exhibit substantial local metrology errors or incorrect defect states, obscuring whether the underlying process-dependent response has been faithfully modeled.

To bridge these gaps, we propose FabBench, a fabrication-informed benchmark for systematically evaluating process-dependent geometric evolution and defect formation in the SEM domain. FabBench formulates process-conditioned lithography modeling as the task of mapping mask layouts and continuous process parameters directly to observable SEM-domain geometry and topological defect states. To enable this setting at scale, we construct a benchmark containing 547K physics-grounded, real-SEM-informed mask–process–SEM triplets by coupling lithography simulation with SEM-informed rendering. With continuous control over illumination source, defocus, and dose, FabBench provides structured in-distribution and out-of-distribution splits across mask topologies, process conditions, and source configurations, enabling reproducible evaluation of parametric generalization. FabBench further introduces a hierarchical, metrology-aware evaluation protocol spanning SEM image fidelity, contour geometry, localized dimensional accuracy, and topological defect consistency. By explicitly measuring fine-grained geometric deviations and defect-state errors, FabBench reveals fabrication-critical failures that are often obscured by conventional global similarity metrics. Together, these components establish a unified testbed for evaluating whether models capture geometric responses and defect transitions of lithography models.

Across diverse baselines, FabBench reveals a consistent performance trade-off: generative models achieve higher visual fidelity, whereas deterministic and operator-based models exhibit more stable local metrology under process variation. Crucially, high contour overlap can still coincide with substantial CD errors and incorrect defect states, demonstrating that global geometric agreement does not necessarily imply fabrication-relevant accuracy. Evaluation on held-out real-wafer data further shows that the relative performance trends observed on FabBench are consistent with those measured on real SEM observations, supporting FabBench as a scalable proxy for identifying metrology- and defect-related failures and for guiding the development of process-conditioned lithography models.

2 Related Work

Lithography Datasets and Benchmarks. Existing lithography benchmarks Banerjee et al. (2013); Liu et al. (2019); Yang et al. (2022b); Ye et al. (2019); Wang et al. (2022); Zheng et al. (2023) predominantly rely on physics-based simulations, representing fabrication outcomes as resist contours or intensity maps. While providing scalable and controllable supervision, these idealized representations remain disconnected from the physical SEM topographies used in actual fabrication inspection and metrology. In parallel, real-wafer SEM datasets focus on defect understanding, such as classification, segmentation, and multimodal interpretation Qiao et al. (2024); Jin et al. (2024); Jiang et al. (2024); Jin et al. (2026). Although grounded in real observations, these studies focus on isolated defect classification and segmentation, rather than process-conditioned lithography modeling that jointly accounts for geometric fidelity and defect status. As a result, process-conditioned geometric modeling and real-world defect status prediction remain fundamentally decoupled. FabBench bridges this divide by providing physics-grounded SEM triplets under controlled process conditions, unifying process-dependent geometric evolution and defect formation within a single benchmark.

Process-Conditioned Lithography Modeling. Recent computational lithography models have advanced beyond nominal-condition prediction toward modeling process-dependent manufacturing responses. Existing approaches encompass deterministic architectures, physics-informed operators, and conditional generative models Wang et al. (2025); He et al. (2025); Jin et al. (2025). Diffusion-based frameworks have been explored for process-conditioned lithography optimization Chen et al. (2024), while state-space architectures such as LithoMamba enhance long-range dependencies for high-fidelity lithography simulation He et al. (2026). More recently, LithoDreamer Jiang et al. (2026) extends simulation from isolated steps to multi-stage physical evolution. Despite these advances, evaluations of SEM-domain predictions largely follow standard vision or contour overlap metrics inherited from resist modeling. Consequently, it remains largely unexamined whether models maintain localized geometric precision and defect-level fidelity under process shifts. To address this, FabBench establishes a hierarchical, metrology-aware protocol that systematically benchmarks localized geometric precision and topological defect fidelity across continuous process shifts.

3 FabBench

3.1 Benchmark Task

Prediction task. FabBench evaluates process-conditioned lithography models that predict SEM-domain images—capturing both fine-grained geometric evolution and topological defect formation from a mask layout and controllable process conditions. We denote the mask layout as M and the process condition as $p = (S, F, E)$, where S is the illumination source, F is the defocus, and E is the exposure dose. These variables interact non-linearly: S dictates the overall optical transfer characteristics for a pattern family, while F and E perturb local edge placements and linewidths, ultimately driving topological defect formation. Formally, given a condition tuple (M, S, F, E), a predictive model g_ϕ must map these inputs to the corresponding SEM-domain image:

$$\hat{I}_{\text{SEM}} = g_\phi(M, S, F, E). \tag{1}$$

Evaluation protocol. FabBench establishes a hierarchical, metrology-aware evaluation protocol that assesses lithography models across multiple levels of granularity. Beyond generic image fidelity and coarse contour overlap, the protocol rigorously evaluates whether predictions preserve the local geometric accuracy and topological defect behaviors critical for fabrication-side metrology.

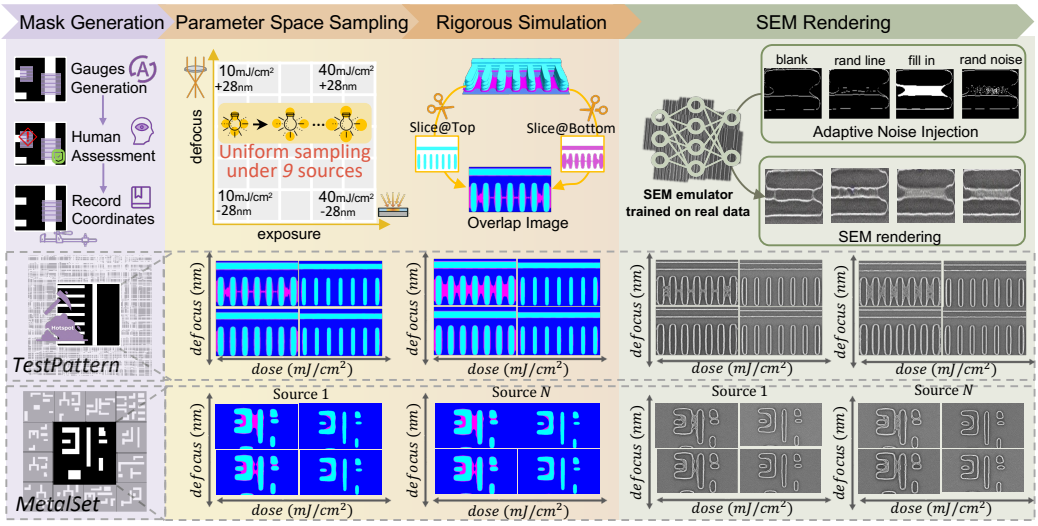


Figure 2: **Physics-grounded SEM Data curation pipeline. Top:** By coupling rigorous physical simulation with real-SEM-informed rendering, FabBench captures both continuous process physics and realistic imaging nuances. **Bottom:** Qualitative transitions from intermediate simulated profiles to final rendered SEM images across diverse focus-exposure conditions and source setups.

3.2 Data Curation Pipeline

Lithography datasets paired with precise, dense process annotations are indispensable for benchmarking process-conditioned prediction models. However, authentic wafer SEM images and their associated manufacturing recipes are proprietary and practically impossible to release at scale. To overcome this fundamental bottleneck, as illustrated in Fig. 2, FabBench introduces a physics-grounded pipeline that synthesizes scalable, high-fidelity SEM targets with controlled process annotations.

Mask pattern generation & Gauge definition. FabBench leverages two complementary mask sources to balance structural difficulty and layout realism. First, we construct a stress-test pattern library focusing on failure-prone critical regions by systematically varying geometric attributes like feature width and spacing. We further sample 500 mask patterns from LithoBench Zheng et al. (2023) to introduce layouts with real-world design statistics. For local metrology evaluation, we establish measurement gauges via a semi-automated protocol combining rule-based heuristics with manual verification. Unlike global contour metrics, that often obscure localized errors, these gauge-based measurements serve as local anchors, ensuring that predictions preserve fabrication-critical geometries. Detailed gauge-generation protocols are provided in Appendix E.1.

Parameter space sampling & Rigorous simulation. To capture representative lithography process variations, we systematically vary three controllable process variables: source, defocus, and exposure. For source ($\mathcal{S}$), we adopt quadrupole configurations, a standard source illumination family with four bright regions in the optical pupil, and sweep the normalized pole radius over the range $[0.17, 0.33]$. Defocus and exposure are sampled uniformly within $[-28, 28]$ nm and $[10, 40]$ mJ/cm^2.

Given a mask-process tuple (M, S, F, E), we employ a rigorous lithography simulator to generate the corresponding 3D resist profile. The simulator integrates physically grounded optical, exposure, and resist-development modules (details in Appendix A.1), yielding process-consistent profiles that serve as the physical foundation for the subsequent SEM rendering.

SEM rendering. To map simulated 3D resist profiles to realistic SEM images, we deploy a generative SEM emulator. This emulator provides scalable, physics-consistent SEM-domain observations paired with strictly controlled process annotations, closing the gap between idealized resist profiles and observable metrology. Built upon ControlNet Zhang et al. (2023a), the emulator is trained directly on real fabrication SEM data. To capture the characteristic SEM "white-edge" effect—driven by secondary electron emission—we condition the renderer on cross-sectional contours extracted from both the top and bottom interfaces of the simulated resist profile. This dual-layer representa-

Table 1: **FabBench Dataset Composition.** Shaded cells indicate held-out distributions relative to the ID training split. †The ID split contains 98,851 valid samples after removing invalid simulations.

Split	Evaluation Objective	Masks (M)	Sources (S)	(F, E) coverage	Samples
ID	In-distribution prediction	111	9 Quadrupole	11 × 9 grid	98,851†
Mask-OOD	Generalize to unseen masks	500	9 Quadrupole	11 × 9 grid	445,500
Process-OOD	Extrapolate to extreme (F, E)	111	1 Quadrupole	17 conditions	1,887
Source-OOD	Test unseen source family	111	1 Quasar	10 conditions	1,110
Total		611	–	–	547,348

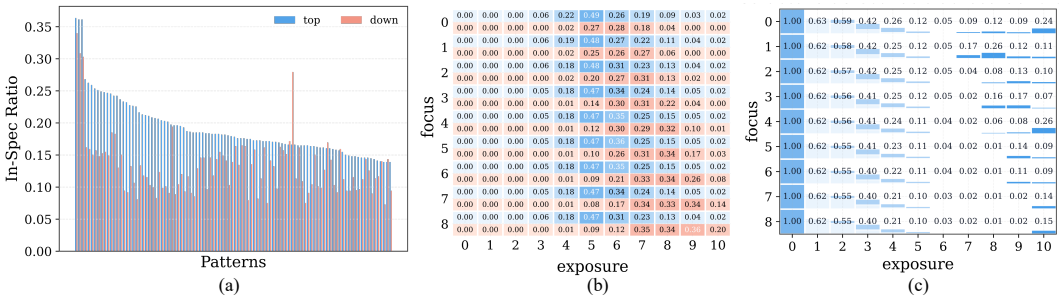


Figure 3: **Data analysis of FabBench.** Spanning diverse focus-exposure conditions with depth-dependent (top/bottom) metrology and varied defect types, FabBench provides a rigorous testbed for diagnosing process-induced geometric degradation and defect formation.

tion captures the 3D sidewall slope, which dictates the intensity near printed edges. Furthermore, as shown in Fig. 2, we apply adaptive noise injection to these contour maps using four distinct strategies to simulate diverse resist-profile conditions and topological states.

3.3 Dataset Analysis

Dataset splits and design rationale. As detailed in Table 1, FabBench comprises 547k structured samples deliberately partitioned to decouple distinct physical shifts. For in-distribution (ID) evaluation, the training, validation, and test sets are partitioned at the mask-process pair level; while sharing individual masks and process grid points, their specific combinations remain strictly mutually exclusive. To diagnose generalization limits, three out-of-distribution (OOD) splits are curated around distinct physical axes. Specifically, Mask-OOD reserves 500 held-out layout geometries to evaluate structural generalization under familiar process distributions. In contrast, Process-OOD and Source-OOD reuse masks from the ID pool while shifting a single physical condition at a time. For Process-OOD, we sample 17 extreme focus-exposure conditions at the usable process-window boundary; this stress-tests extrapolation limits while avoiding complete pattern wash-out into trivial non-imaging failures. Finally, Source-OOD serves as a targeted optical-transfer probe, pairing ID layouts with an unseen quasar illumination family across representative focus-exposure conditions.

Dataset statistics analysis. We further verify whether FabBench faithfully captures the physical dependencies between controllable process conditions, continuous geometric deformations, and discrete defect formations. First, we quantify depth-dependent geometric variations by extracting top and bottom contours from the SEM-domain targets. As shown in Fig. 3(a), the ratio of on-spec critical dimensions within the process window varies substantially across layouts (from over 35% on robust masks to below 10% on sensitive ones). The discrepancy between top and bottom CDs directly quantifies 3D profile deformation, demonstrating that FabBench spans diverse mask-level process tolerances. Next, Fig. 3(b) demonstrates how geometric yield is explicitly controlled by process conditions across the focus-exposure grid: the within-tolerance ratio peaks at the nominal process center and systematically deteriorates toward peripheral defocus-exposure regimes. Finally, the defect map in Fig. 3(c) establishes the topological consequence of these geometric excursions: as severe continuous CD degradation pushes profiles beyond critical printability thresholds, it trig-

gers catastrophic discrete defects (e.g., pinching and bridging). Together, these analyses confirm that FabBench embodies physically coherent variations spanning continuous 3D profile distortions to discrete topological failures across the entire process space. More detailed statistics regarding data distributions for OOD splits are provided in Appendix B.

3.4 Evaluation Metrics

Image quality. We first employ generic image metrics to evaluate the basic SEM-domain synthesis capability of the models. To measure whether the predicted SEM image recover the overall appearance and distributional characteristics of the target, we report paired image fidelity metrics, including PSNR and LPIPS, as well as dataset-level similarity measured by FID.

Geometric fidelity. To evaluate structural preservation, we utilize a fine-tuned Segment Anything Model (SAM) to extract top and bottom contours from the SEM images. Using these dual-layer contours, we compute the Dice coefficient and Boundary IoU (B-IoU) to assess global shape deformation and boundary-specific errors such as corner rounding. Going beyond coarse overlap, we evaluate local metrology. Specifically, we compute Critical-Dimension (CD) errors along predefined gauge directions to measure linewidth accuracy, and Edge Placement Error (EPE) along mask boundaries to quantify fine-grained local edge shifts. To ensure a consistent evaluation, the same extraction and metrology pipeline is applied to both the ground-truth targets and model predictions. The reliability of our SAM-based extractor is validated on a held-out, expert-calibrated test set, achieving robust performance: 96.13% B-IoU, $0.45\,\mathrm{nm}$ CD error, 98.27% defect-status F1, and 96.70% defect-type F1 (see Appendix D for full details).

Defect-level sensitivity. This metric group evaluates whether models preserve process-induced topological failures rather than hallucinating nominal printed patterns. Based on the extracted contours and the corresponding masks, we assign rule-based defect labels to evaluate binary defect status and specific failure categories. We report *Defect-status F1* for binary classification and multi-label *Defect-type F1* over five non-exclusive categories: bridge, pinch, scumming, missing, and washout. Type F1 is macro/micro: macro-F1 averages the five per-class scores, micro-F1 pools all sample–class decisions, and zero-support classes contribute zero to the macro average.

Efficiency. Finally, we report turnaround time (TAT) under a standardized hardware setting to measure the practical cost of large-scale process-window exploration. Detailed mathematical formulations of all metrics and defect rules are provided in the AppendixE.1.

4 Experiments

4.1 Benchmark Models

To provide a comprehensive evaluation, FabBench benchmarks three representative model families:

- **Deterministic prediction:** This group includes standard CNN- and Transformer-based image-to-image architectures. To incorporate lithography process conditions, these models follow the conditional encoding scheme used in LithoSim He et al. (2025), where source, defocus, and exposure parameters are embedded into the backbone through lithography-aware cross-attention.
- **Neural operator learning:** We evaluate RFNO and MFNO Yang et al. (2022a); Li et al. (2024) to assess whether spectral inductive biases are beneficial for SEM-domain process-conditioned lithography models. They share the same process-conditioning interface as the deterministic baselines. We also attempted to reproduce CFNO Yang & Ren (2023), but observed unstable convergence on FabBench; we therefore report its reproduction details in Appendix H and exclude it from the main quantitative comparison.
- **Conditional generation:** This group includes GAN- and diffusion-based models. For Pix2Pix Isola et al. (2016), process parameters are spatially expanded and concatenated with the input mask as early fusion. For Stable Diffusion, we evaluate multiple conditioning mechanisms, including intermediate AdaIN modulation and a lithography-adapter formulation.

All models take the same input tuple (M, S, F, E) and predict the corresponding rendered SEM image. They are trained and evaluated on the same FabBench splits using the same metrology-aware evaluation protocol. Detailed architectures and training settings are provided in Appendix H.

Table 2: Main benchmark results on ID and Mask-OOD. Geometric metrics average both contours; multi-label Type F1 is macro/micro. Best and second-best use macro-F1 for the dual-value col.

Split	Method	Image Quality			Geometric Fidelity				Defect Sensitivity		Efficiency
		PSNR ↑	LPIPS ↓	FID ↓	Dice ↑	B-IoU ↑	EPE (nm) ↓	CD (nm) ↓	Status F1 ↑	Type F1 (Ma/Mi) ↑	TAT ↓
ID	CNN	24.439	0.658	163.17	0.956	0.763	2.70	22.66	0.709	.398/.418	0.029
	Trans	24.417	0.658	162.13	0.939	0.731	3.75	31.64	0.720	.438/.471	0.028
	RFNO	21.847	0.679	166.14	0.920	0.660	4.51	40.65	0.740	.340/.346	0.013
	MFNO	24.441	0.658	165.26	0.959	0.768	2.53	23.09	0.679	.415/.415	0.008
	Pix2Pix	22.055	0.180	70.53	0.958	0.766	2.34	15.05	0.730	.405/.502	0.013
	SD-AdaIN	19.127	0.208	24.14	0.938	0.665	3.80	32.12	0.733	.443/.513	4.453
	SD-Adapter	20.457	0.175	15.41	0.962	0.770	2.40	20.93	0.744	.469/.523	4.447
Mask OOD	CNN	23.961	0.718	144.64	0.954	0.737	1.98	24.22	0.837	.416/.612	0.029
	Trans	23.072	0.743	147.83	0.922	0.661	2.65	41.55	0.608	.408/.592	0.028
	RFNO	23.209	0.735	146.30	0.940	0.667	3.35	21.33	0.761	.381/.572	0.013
	MFNO	24.348	0.715	143.77	0.962	0.752	2.52	19.05	0.858	.430/.666	0.008
	Pix2Pix	23.198	0.193	69.09	0.966	0.783	1.98	11.17	0.863	.492/.764	0.013
	SD-AdaIN	20.554	0.240	42.90	0.921	0.663	3.91	112.93	0.617	.288/.403	4.453
	SD-Adapter	20.806	0.197	18.98	0.968	0.784	1.90	26.98	0.869	.431/.637	4.447

4.2 Benchmark Results

Model-family behavior. As shown in Tables 2 and 3, FabBench reveals distinct complementary strengths and failure modes across different model families. Deterministic regressors and operator-based models are computationally efficient and provide strong baselines for direct mask-process-to-rendered SEM prediction. Although they are not always strongest in LPIPS or FID, they exhibit consistently more stable local-metrology behavior across both ID and OOD splits. For instance, CNN and MFNO keep CD errors within relatively narrow ranges across four splits (10.78–24.22 for CNN and 9.29–23.09 for MFNO), whereas SD-Adapter rises to 54.97 under Process-OOD and 34.04 under Source-OOD. This clearly suggests stronger local-metrology robustness for deterministic and operator-based baselines under drastic physical-condition shifts.

Generative baselines exhibit a distinct trade-off between visual realism and metrological precision. As shown in Fig. 4, these models produce significantly more realistic SEM textures than conventional deterministic predictors. On ID and Mask-OOD splits, generative variants are highly competitive: SD-Adapter achieves the leading FID, B-IoU, and ID multi-label macro-F1, while Pix2Pix yields the lowest CD error and the best Mask-OOD macro/micro-F1. This clearly indicates that generative architectures excel at SEM-domain prediction when process conditioning remains largely interpolative. However, in the Process-OOD split, the CD error of Pix2Pix rises from $15.05\,\mathrm{nm}$ (ID) to $22.19\,\mathrm{nm}$, whereas SD-Adapter's error surges to $54.97\,\mathrm{nm}$. This stark disparity suggests that while current SD-based models effectively capture macro distributional statistics, their latent-space conditioning lacks the robust control required to model linewidth changes under extreme process extrapolations, as discussed in Appendix H.3. In contrast, Pix2Pix's direct image-to-image translation preserves much finer-grained linewidth control, offering a more stable balance between generative realism and geometric fidelity when facing unseen physical conditions.

OOD performance and failure modes. The disentangled OOD splits in FabBench reveal that mask transfer, process extrapolation, and source adaptation correspond to distinct generalization axes. On the unseen *MetalSet* masks, most models retain strong performance under in-distribution process conditions. Pix2Pix achieves the lowest CD error and the best multi-label defect-type macro/micro-F1, while SD-Adapter obtains the strongest Dice and Boundary IoU. Together with the dataset analysis in Appendix B, which shows that *MetalSet* is more process-stable than the sensitive *TestPattern* masks, these results suggest that training on optically challenging test patterns provides useful coverage of local proximity primitives and can transfer to less sensitive mask distributions. Notably, the Transformer baseline yields a higher CD error on *MetalSet* (41.55 nm) than on the ID *TestPattern* split (31.64 nm), suggesting that mask transfer is sensitive to the representation of layout geometry. One possible explanation is that tokenized mask representations capture global context but may lose fine local structures that are critical for linewidth metrology under unseen mask topologies.

Process-OOD exposes a sharp split between image quality and local metrology. SD-Adapter gives the best FID, but its CD error remains high (54.97 nm), whereas MFNO and CNN achieve much

Table 3: Main benchmark results on Process-/Source-OOD. Geometric metrics, Status F1, and multi-label Type F1 give the mean ± bootstrap 95% CI half-width; Type F1 is macro/micro. Best and second-best use macro-F1 for the dual-value column.

Split	Method	Image Quality			Geometric Fidelity				Defect Sensitivity	
		PSNR ↑	LPIPS ↓	FID ↓	Dice ↑	B-IoU ↑	EPE (nm) ↓	CD (nm) ↓	Status F1 ↑	Type F1 (Ma/Mi) ↑
Process OOD	CNN	24.296	0.721	243.410	0.922 ± 0.023	0.771 ± 0.042	3.08 ± 0.78	10.78 ± 8.15	0.354 ± 0.086	.144 ± .044 / .110 ± .035
	Trans	24.223	0.719	233.700	0.917 ± 0.024	0.758 ± 0.044	3.31 ± 0.81	17.82 ± 13.65	0.340 ± 0.087	.183 ± .042 / .117 ± .034
	RFNO	24.252	0.724	243.812	0.893 ± 0.028	0.710 ± 0.048	4.31 ± 0.93	15.58 ± 12.09	0.354 ± 0.085	.209 ± .039 / .161 ± .036
	MFNO	24.223	0.723	252.885	0.914 ± 0.024	0.742 ± 0.044	3.50 ± 0.83	9.29 ± 5.64	0.355 ± 0.085	.183 ± .042 / .129 ± .035
	Pix2Pix	22.361	0.267	194.512	0.892 ± 0.027	0.686 ± 0.048	4.23 ± 0.85	22.19 ± 15.87	0.347 ± 0.092	.183 ± .039 / .143 ± .036
	SD-AdaIN	20.512	0.216	73.308	0.916 ± 0.028	0.682 ± 0.065	4.58 ± 1.24	52.94 ± 32.47	0.353 ± 0.120	.222 ± .068 / .138 ± .041
	SD-Adapter	20.467	0.212	57.891	0.916 ± 0.024	0.715 ± 0.048	4.20 ± 0.97	54.97 ± 27.90	0.356 ± 0.087	.235 ± .037 / .135 ± .035
Source OOD	CNN	19.767	0.631	181.428	0.927 ± 0.014	0.627 ± 0.044	2.94 ± 0.52	22.66 ± 11.86	0.451 ± 0.143	.121 ± .040 / .137 ± .038
	Trans	19.644	0.632	182.806	0.929 ± 0.012	0.623 ± 0.043	2.93 ± 0.49	24.22 ± 12.26	0.464 ± 0.149	.103 ± .045 / .130 ± .041
	RFNO	17.987	0.667	222.271	0.892 ± 0.018	0.527 ± 0.047	4.21 ± 0.70	25.47 ± 11.89	0.479 ± 0.144	.051 ± .040 / .055 ± .032
	MFNO	18.144	0.660	218.792	0.894 ± 0.018	0.540 ± 0.047	3.96 ± 0.65	20.17 ± 9.68	0.479 ± 0.143	.068 ± .040 / .074 ± .034
	Pix2Pix	19.202	0.193	80.901	0.926 ± 0.015	0.624 ± 0.041	3.21 ± 0.57	24.49 ± 12.31	0.420 ± 0.135	.110 ± .041 / .157 ± .045
	SD-AdaIN	17.341	0.222	69.272	0.886 ± 0.023	0.525 ± 0.039	4.08 ± 0.61	27.83 ± 13.32	0.432 ± 0.143	.091 ± .046 / .097 ± .039
	SD-Adapter	17.508	0.212	64.640	0.885 ± 0.027	0.554 ± 0.043	4.28 ± 0.84	34.04 ± 18.02	0.410 ± 0.132	.097 ± .042 / .104 ± .040

lower CD errors (9.29 nm and 10.78 nm). By comparison, the overlap metrics have much narrower intervals, such as Dice scores of 0.916 for SD-Adapter and 0.922 for CNN. Therefore, the main Process-OOD failure mode shows a loss of process-controlled linewidth accuracy. Source-OOD further tests adaptation to a changed illumination source family, which alters the optical transfer characteristics rather than only the scalar process variables. As illustrated in Fig. 4(d), most models still preserve coarse printed shapes, but multi-label defect-type F1 remains low for all methods (at most 0.121 macro-F1 and 0.157 micro-F1). This suggests that topology-level failure prediction remains difficult under unseen illumination even when coarse geometry remains plausible.

Validation on Real-wafer Data. We further evaluate the benchmark models on held-out wafer SEM data to examine whether conclusions drawn from FabBench transfer to real fabrication observations (Table 4; Appendix G). The resulting model rankings remain exceptionally consistent with those on FabBench ($\rho = 0.829$–0.986; Table 10), indicating that the benchmark faithfully preserves meaningful relative differences between model families beyond its SEM-emulated setting. More importantly, real-wafer evaluation reinforces the critical discrepancy revealed by FabBench: models with similar global contour agreement can exhibit substantially different boundary and local-metrology behavior, particularly on challenging dense layouts (Table 11). Additional sensitivity analyses show that these observations are stable to defect-label thresholds (Table 13), while direct comparisons with LithoSim further confirm that similar overlap scores can conceal markedly different CD accuracy (Table 14). Together, these empirical results support FabBench as a practically relevant proxy for diagnosing fabrication-oriented model behavior, while highlighting the need to evaluate lithography models beyond standard visual and contour-level metrics.

Table 4: Held-out wafer-SEM evaluation.

Model	CD ↓	EPE ↓	B-IoU ↑	FID ↓	Status F1 ↑
CNN	24.40	3.19	0.689	142.32	0.771
Transformer	30.75	3.76	0.648	149.72	0.674
RFNO	19.80	3.91	0.660	147.21	0.802
MFNO	15.48	3.38	0.670	145.48	0.818
Pix2Pix	**14.68**	3.25	0.682	85.48	0.849
SD-Adapter	19.84	**3.12**	**0.691**	**38.91**	**0.867**

Validation of SEM emulator. To evaluate the SEM emulator, we input contours extracted from a held-out set of real wafer SEM images and compare the synthesized outputs against their real counterparts. This validation follows the contour extraction and evaluation protocols used in FabBench. The rendered images show strong consistency with real SEM observations, achieving a low CD error of 2.12 nm and a high defect-type F1 of 93.1%. To verify cross-domain robustness, we further test the emulator's ability to adapt to a neighboring inspection process. Compared to the 10 epochs required for initial training, fine-tuning on a 25% data subset rapidly restores high fidelity in just 3.75 effective epochs, proving its adaptability and sample efficiency for new SEM domains. Adaptive noise injection strategies and more comprehensive quantitative results are detailed in Appendix C.

Hierarchical metric analysis. FabBench shows that metrology-aware evaluation is naturally organized from coarse to fine. At the coarse level, Dice measures whether the predicted contour region overlaps the target contour, while Boundary IoU further probes contour-level localization. Across

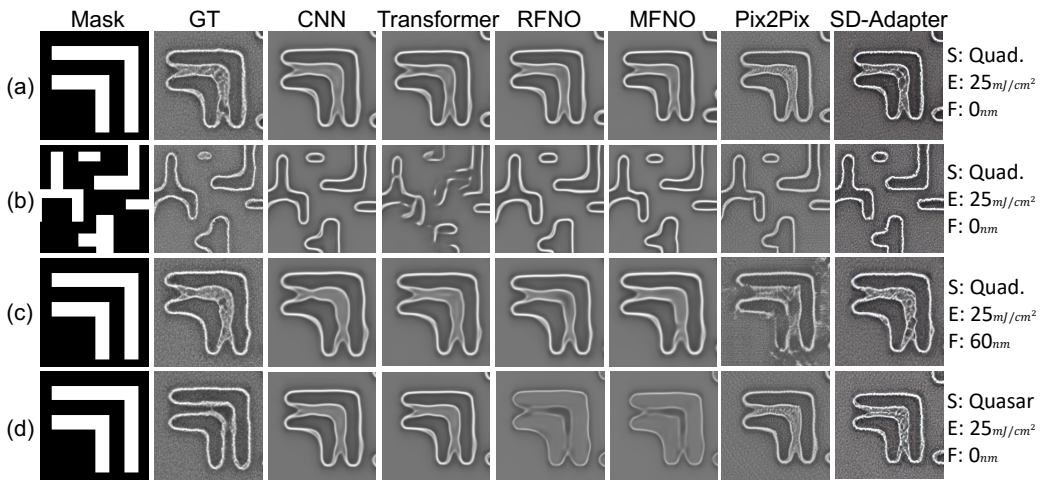


Figure 4: **Qualitative comparison across representative settings.** We compare predictions under ID, Mask-OOD, Process-OOD, and Source-OOD conditions, covering unseen masks, focus–exposure shifts, and illumination families. This reveal increasingly pronounced geometric and defect degradation as models move from in-distribution to more challenging out-of-distribution settings.

most splits, most models retain high coarse-overlap scores; for example, under Process-OOD, CNN and SD-Adapter reach Dice scores of 0.922 and 0.916, respectively. However, as shown in metrics sensitivity analysis in Appendix E.2, coarse contour accuracy does not guarantee correct local metrology. On Mask-OOD, SD-Adapter achieves the best Dice, Boundary-IoU, and EPE, yet its CD error is substantially higher than Pix2Pix and MFNO. This occurs because EPE measures edge location, whereas CD measures linewidth. A model may localize the overall contour reasonably well, yet still predict an incorrect spacing between opposite edges.

Among all metrics, defect evaluation probes discrete topological failures. Unlike continuous linewidth prediction, faithfully capturing topological defect formation requires models to resolve process-induced connectivity changes and depth-dependent contour interactions. Multi-label Type F1 remains low across model families under Process-OOD and Source-OOD: the best macro-F1 values are 0.235 and 0.121, respectively. This difficulty partly reflects the sparse and process-sensitive occurrence of bridge, pinch, scumming, missing, and washout events, but it also demonstrates that current models fail to reliably capture process-driven defect formation. We therefore treat defect formation accuracy as a strict diagnostic metric and a crucial open challenge.

Discussion. Overall, FabBench demonstrates that process-conditioned lithography models cannot be properly evaluated by conventional visual fidelity or single leaderboard scores. As a fabrication-informed benchmark, FabBench reveals a fundamental trade-off across distinct model families: while generative approaches excel at synthesizing realistic SEM textures, deterministic and operator-based models exhibit far superior metrological stability under drastic physical condition shifts. Crucially, our multi-axis evaluation empirically exposes that high contour overlap frequently conceals severe CD errors and critical failures in capturing topological defect formation. FabBench bridges controllable process conditions with observable SEM metrology. Validated by real-wafer data, FabBench bridges controllable process conditions into the SEM domain, providing a reliable diagnostic testbed for process-conditioned lithography modeling.

5 Conclusion

We present FabBench, the first fabrication-informed benchmark for systematically evaluating process-conditioned lithography models in the SEM domain. By assessing geometric evolution and topological defect formation, FabBench reveals fabrication-critical failures that conventional visual and contour-overlap metrics often overlook. Supported by held-out real-wafer observations, FabBench provides a reliable testbed for evaluating whether lithography models faithfully reproduce process-dependent geometric responses and defect transitions.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. FabBench studies process-conditioned lithography models using physics-based simulation, SEM-emulated benchmark targets, and a held-out wafer-SEM validation. The released rendered images are synthetic or emulated observations and should not be presented as verified measurements from a specific fabrication line. FabBench and its baselines may inherit biases and limitations from the lithography simulator, the frozen SEM emulator, the SAM-based contour extractor, and the selected mask and process distributions. Raw proprietary SEM images, fabrication recipes, equipment calibration parameters, and sensitive industrial layouts are excluded from release. Use and redistribution of benchmark data, pretrained or frozen models, and third-party assets remain subject to their original licenses and to applicable privacy, confidentiality, institutional, industrial, and export-control requirements.

REPRODUCIBILITY STATEMENT

Section 3.2 describes the mask generation, process-space sampling, rigorous lithography simulation, and SEM rendering pipeline. Section 4.2 specifies the evaluated model families, ID/OOD protocol, reported metrics, and additional wafer-SEM, dense-layout, and multi-label evaluations. Appendix A.1 details the lithography and resist models; Appendix B documents dataset composition and process sensitivity; Appendix C describes SEM-emulator training and validation; Appendix D describes contour extraction; and Appendix E.1 defines the evaluation metrics and defect rules. Appendix H provides model architectures, conditioning mechanisms, training settings, and inference details. The benchmark code and releaseable data will be made available upon acceptance, subject to the stated restrictions on proprietary fabrication data.

AI USE STATEMENT

We used large language models to polish the manuscript, including rephrasing sentences, correcting grammar, and improving readability and flow. The authors reviewed and edited the AI-assisted text to ensure that it accurately reflects the methods, experiments, and results, and take full responsibility for the content of the manuscript.

REFERENCES

S. Banerjee, Z. Li, and S. R. Nassif. Iccad-2013 cad contest in mask optimization and benchmark suite. In *2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 271–274, 2013. ISBN 1558-2434.

Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013. ISBN 148310320X.

Guojin Chen, Zixiao Wang, Bei Yu, David Z Pan, and Martin DF Wong. Ultrafast source mask optimization via conditional discrete diffusion. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 43(7):2140–2150, 2024.

Hongquan He, Zhen Wang, Jingya Wang, Tao Wu, Xuming He, Bei Yu, Jingyi Yu, and Hao Geng. Lithosim: A large, holistic lithography simulation benchmark for ai-driven semiconductor manufacturing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Xinyu He, Daohui Wang, Shujing Lyu, Pourya Shamsolmoali, Jiwei Shen, and Yue Lu. Lithomamba: High-fidelity lithography simulation with state space models. In *2026 Design, Automation & Test in Europe Conference (DATE)*, pp. 1–6. IEEE, 2026.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2016.

Yuqi Jiang, Xudong Lu, Qian Jin, Qi Sun, Hanming Wu, and Cheng Zhuo. Fabgpt: An efficient large multimodal model for complex wafer defect knowledge queries. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pp. 1–8, 2024.

Yuqi Jiang, Yumeng Liu, Zimu Li, Jinyuan Deng, Qian Jin, Yucheng Cui, Yu Li, Xunzhao Yin, Qi Sun, and Cheng Zhuo. Lithodreamer: A physics-informed world model for multi-stage computational lithography. *arXiv preprint arXiv:2606.26713*, 2026.

Qian Jin, Yuqi Jiang, Xudong Lu, Yumeng Liu, Yining Chen, Dawei Gao, Qi Sun, and Cheng Zhuo. Sem-clip: Precise few-shot learning for nanoscale defect detection in scanning electron microscope image. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pp. 1–8, 2024.

Qian Jin, Yumeng Liu, Yuqi Jiang, Qi Sun, and Cheng Zhuo. Unitho: A unified multi-task framework for computational lithography. *arXiv preprint arXiv:2511.10255*, 2025.

Qian Jin, Ruidong Li, Yuqi Jiang, Yumeng Liu, Xudong Lu, Yining Chen, Dawei Gao, Qi Sun, and Cheng Zhuo. Sem-clip 2.0: Precise zero-/few-shot learning for nanoscale defect detection in sem image. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2026.

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.

Wen-Hao Liu, Stefanus Mantik, Wing-Kai Chow, Yixiao Ding, Amin Farshidi, and Gracieli Posser. Ispd 2019 initial detailed routing contest and benchmark with advanced routing rules. In *Proceedings of the 2019 international symposium on physical design*, pp. 147–151, 2019.

Yibo Qiao, Zhouzhouzhou Mei, Yuening Luo, and Yining Chen. Deepsem-net: Enhancing sem defect analysis in semiconductor manufacturing with a dual-branch cnn-transformer architecture. *Computers & Industrial Engineering*, 193:110301, 2024.

Q. Wang, X. Gao, Y. Lin, R. Wang, and R. Huang. Deepeb: A neural partial differential equation solver for post exposure baking simulation in lithography. In *2022 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pp. 1–9, 2022. ISBN 1558-2434.

Zhen Wang, Hongquan He, Tao Wu, Xuming He, Qi Sun, Cheng Zhuo, Bei Yu, Jingyi Yu, and Hao Geng. Lmlitho: A large vision model-driven lithography simulation framework. *2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pp. 1–9, 2025.

Haoyu Yang and Haoxing Ren. Enabling scalable ai computational lithography with physics-inspired models. In *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, pp. 715–720, 2023.

Haoyu Yang, Zongyi Li, Kumara Sastry, Saumyadip Mukhopadhyay, Mark Kilgard, Anima Anandkumar, Brucek Khailany, Vivek Singh, and Haoxing Ren. Generic lithography modeling with dual-band optics-inspired neural networks. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pp. 973–978, 2022a.

Haoyu Yang, Zongyi Li, Kumara Sastry, Saumyadip Mukhopadhyay, Mark Kilgard, Anima Anandkumar, Brucek Khailany, Vivek Singh, and Haoxing Ren. Generic lithography modeling with dual-band optics-inspired neural networks. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pp. 973–978, 2022b.

W. Ye, M. B. Alawieh, Y. Lin, and D. Z. Pan. Lithogan: End-to-end lithography modeling with generative adversarial networks. In *2019 56th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6, 2019. ISBN 0738-100X.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023a.

Shuang Zhang, Libin Zhang, Chenghong Ding, Lei Wang, Hongru Zhang, Ming Ding, Shengrui Zhang, Weijie Shi, and Yayi Wei. Dtco optimizes critical path nets to improve chip performance with timing-aware opc in deep ultraviolet lithography. *Applied optics*, 62(27):7216–7225, 2023b.

Su Zheng, Haoyu Yang, Binwu Zhu, Bei Yu, and Martin D. F. Wong. Lithobench: Benchmarking ai computational lithography for semiconductor manufacturing. In *Neural Information Processing Systems*, 2023.

CONTENTS

1 Introduction **1**

2 Related Work **3**

3 FabBench **3**

3.1 Benchmark Task 3

3.2 Data Curation Pipeline 4

3.3 Dataset Analysis 5

3.4 Evaluation Metrics 6

4 Experiments **6**

4.1 Benchmark Models 6

4.2 Benchmark Results 7

5 Conclusion **9**

A Lithography and physical simulation details **14**

A.1 Rigorous Lithography Simulation 14

A.2 Imaging model 14

A.2.1 Resist model 15

B Dataset Analysis **16**

C SEM emulator details and validation **23**

D Fine-tune SAM for Auto-evaluation **26**

E Evaluation metric details **27**

E.1 Metrics definition 27

E.2 Metrics analysis 29

F External wafer-SEM CD check **30**

G Additional experiments and analyses from the author response **32**

G.1 Evaluation on held-out wafer SEM 32

G.2 Dense-layout stress test 32

G.3 Multi-label defect evaluation 32

G.4 Sensitivity of the threshold-dependent defect rule 33

G.5 Additional dataset and external-benchmark checks 33

H Benchmark model architectures and training details **34**

H.1 Deterministic Regression Models 34

H.2 Physics-Informed Neural Operators 34
H.3 Generative Models 35
H.4 Implementation Details 37

I Additional qualitative results **38**

J Broader Impacts **41**

K Limitations and future work **41**
K.1 Dataset Diversity and Scale 41
K.2 SEM-Emulated Targets and Real-Fab Validation 42
K.3 Scope of Lithography Settings and Process Variables 42
K.4 Future Work 43

A LITHOGRAPHY AND PHYSICAL SIMULATION DETAILS

A.1 RIGOROUS LITHOGRAPHY SIMULATION

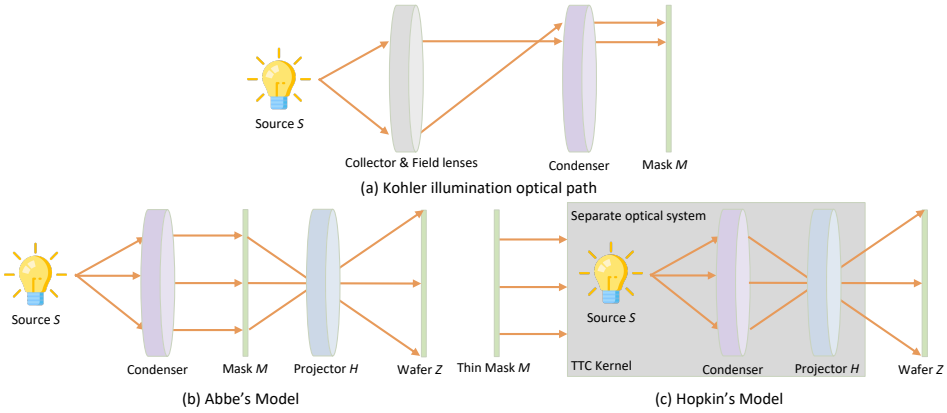


Figure 5: **Illustration of the partially coherent imaging system.** The setup utilizes the Köhler illumination configuration, where the source plane is conjugate to the pupil plane of the objective lens.

A.2 IMAGING MODEL

A lithography imaging system is fundamentally modeled as a partially coherent imaging system, typically employing the Köhler illumination configuration. In this configuration, the effective light source is located at the front focal plane of the condenser lens, and its image is formed at the entrance pupil of the projection objective, ensuring that the mask plane is illuminated by a uniform angular spectrum of plane waves rather than a direct projection of the source geometry Born & Wolf (2013).

The formation of the aerial image $I(x, y)$ is mathematically described by the superposition of coherent diffraction fields generated by each point on the effective source. This interaction between the illumination source distribution S, the mask spectrum M, and the pupil function H of the projection system is formulated as:

$$I(x,y) = \int_{-\infty}^{\infty}\iiint S(f,g)\mathcal{F}(M)(f',g')\mathcal{F}(M)^*(f'',g'') \times H(f+f',g+g')H^*(f+f'',g+g'') \times \exp\left(-j2\pi\big((f'-f'')x+(g'-g'')y\big)\right) \mathrm{d}f\mathrm{d}g\mathrm{d}f'\mathrm{d}g'\mathrm{d}f''\mathrm{d}g'' \quad (2)$$

where $I(x, y)$ is the aerial intensity distribution; $S(f, g)$ denotes the effective source intensity distribution; H represents the amplitude transfer function of the projection system; and $\mathcal{F}(M)$ is the Fourier spectrum of the mask pattern M. The coordinates (x, y) denote the spatial position on the wafer plane, while (f, g), (f', g'), and (f'', g'') represent the normalized frequency-domain coordinates of the source, the mask spectrum, and its conjugate, respectively. Under the assumption of the Köhler illumination configuration, two types of partially coherent imaging models are discussed in the following. The first one is Abbe's model, and the second one is the Hopkins diffraction model, where the Hopkins diffraction model is a simplified and approximate version of Abbe's model.

Abbe's model is based on the physical intuition of source discretization. It decomposes the partially coherent source into a series of mutually incoherent point sources. Each point source generates a plane wave that illuminates the mask. The coherent image formed on the wafer by each point source is calculated independently. The final aerial image $I(x, y)$ is the incoherent summation of the imaging intensities of all these independent point sources:

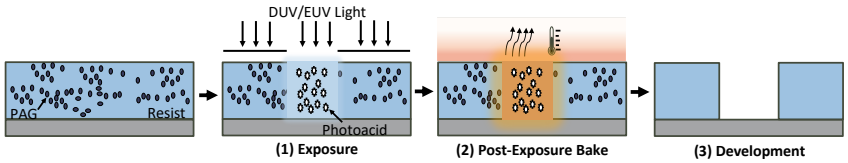


Figure 6: **Explanation of the resist model.** The simulation serves as a bridge transforming the optical aerial image into the physical 3D resist profile via exposure, bake, and development process.

$$I_{Abbe}(x,y) = \iint_{-\infty}^{\infty} S(f,g) \left| \iint_{-\infty}^{\infty} \mathcal{F}(M)(f',g') H(f+f', g+g') e^{-j2\pi(f'x+g'y)} \mathrm{d}f' \mathrm{d}g' \right|^2 \mathrm{d}f \mathrm{d}g \tag{3}$$

To accelerate computation, Hopkins proposed a method to separate the optical system from the mask. Under the thin mask approximation, the Transmission Cross-Coefficients (TCC) are introduced. The TCC matrix is pre-calculated and contains all information regarding the source S and the pupil function H:

$$\mathrm{TCC}(f',g';f'',g'') = \iint_{-\infty}^{\infty} S(f,g) H(f+f', g+g') H^*(f+f'', g+g'') \mathrm{d}f \mathrm{d}g \tag{4}$$

Thus, the Hopkins' approach can be formulated as follows:

$$\begin{aligned} I_{Hopkins}(x,y) = \iiiint_{-\infty}^{\infty} & \mathrm{TCC}(f',g';f'',g'') \mathcal{F}(M)(f',g') \mathcal{F}(M)^*(f'',g'') \\ & \times \exp\left[-j2\pi\left((f'-f'')x + (g'-g'')y\right)\right] \\ & \times \mathrm{d}f' \, \mathrm{d}g' \, \mathrm{d}f'' \, \mathrm{d}g''. \end{aligned} \tag{5}$$

In the mask optimization process, since the TCC remains constant, only the mask portion needs to be updated, resulting in extremely high computational efficiency. However, the TCC matrix in Hopkins' model is typically very large, making direct calculation time-consuming. SOCS is a further approximation of Hopkins' model. It applies singular value decomposition (SVD) to the TCC matrix. Only the top Q eigenvalues and eigenvectors (kernels) are retained. This is equivalent to decomposing a partially coherent imaging system into a weighted sum of a few coherent systems:

$$\mathrm{TCC}(f',g';f'',g'') \approx \sum_{k=1}^{Q} \lambda_k \Phi_k(f',g') \Phi_k^*(f'',g'') \tag{6}$$

where λ_k is the k-th eigenvalue of the TCC matrix, sorted in descending order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_Q$), $\Phi_k(f',g')$ represents the k-th eigenvector, physically corresponding to the transfer function (optical kernel) of the k-th coherent system, Q is the truncation order. Typically, a small number of kernels (e.g., $Q \approx$ 20-60) is used to capture the majority of the aerial image energy.

While Hopkins/SOCS is efficient for thin mask approximation-based optimization, it struggles to accurately model thick-mask effects. To introduce rigorous 3D Mask effects into our simulation and approximate manufacturing reality, we adopt Abbe's model in our data construction pipeline. Specifically, we adopt pFDTD as our Maxwell Solver to rigorously simulate electromagnetic field for each incident angle, thereby accurately capturing pattern shifts, best focus drifts, and non-telecentric effects induced by 3D mask topography.

A.2.1 Resist model

While the imaging model calculates the aerial image intensity $I(x,y)$ projected onto the wafer, the resist model is responsible for simulating how the photoresist material responds to this light energy chemically and physically. Before lithography, the wafer surface is pre-coated with photoresist. Through photochemical reactions, the resist records the aerial image intensity, which is then transformed into a 3D physical profile after the development stage.

In Deep Ultraviolet (DUV, 193nm) and Extreme Ultraviolet (EUV, 13.5nm) lithography, the energy of the light source is insufficient to directly induce the necessary solubility change in the bulk polymer efficiently. To solve this, chemically amplified resists (CAR) are used.

Exposure The concentration of generated acid, $H(x, y)$, is directly related to the aerial image intensity $I(x, y)$ and the exposure time t. Following Dill's model, the acid concentration is modeled as:

$$H(x, y) = 1 - e^{-C \cdot I(x,y) \cdot t}$$

where the initial PAG concentration is normalized to 1.0 and C is the photoreaction rate constant (Dill C parameter).

We instantiate this model using standard ArF chemically amplified resist (CAR) parameters. First, regarding the reaction kinetics: we set the Dill C parameter to $0.0364\,\mathrm{cm}^2/\mathrm{mJ}$, which directly dictates the sensitivity of the acid generation rate in the exponential term. Second, regarding the optical absorption (which modulates the intensity I): the resist's optical properties are defined by a refractive index $n = 1.72$ and an extinction coefficient $k = 0.018$. These correspond to a non-bleachable absorption (Dill B) of $1.155\,\mu\mathrm{m}^{-1}$ and a negligible bleachable absorption (Dill A) of $0.0014\,\mu\mathrm{m}^{-1}$.

Post-Exposure Bake During PEB, a reaction and diffusion process happens. First, the acid attacks the inhibitor (protecting group) on the polymer, changing the concentration of the inhibitor (M). Then the acid ions move through the resist matrix due to thermal energy. Diffusion smoothes out 'standing waves' (interference patterns in the vertical direction) caused by light reflection. Excessive diffusion blurs the latent image, causing critical dimension (CD) loss and limiting resolution. In our model, we set 90s for PEB time.

Development We use the Mack model to map the inhibitor concentration (M) to the dissolution rate (R) of the resist. The Mack model assumes a non-linear 'threshold-like' behavior: areas with high inhibitor concentration (unexposed) dissolve very slowly (R_{min}) and areas with low inhibitor concentration (exposed and deprotected) dissolve very quickly (R_{max}). The dissolution rate $R(M)$ is calculated as:

$$R(M) = R_{max} \frac{(a+1)(1-M)^n}{a + (1-M)^n} + R_{min}$$

Where R_{max} indicates the maximum dissolution rate (fully deprotected resist); R_{min} is the minimum dissolution rate (fully protected resist); n is the dissolution selectivity parameter (contrast), controlling the slope of the transition. Higher n means higher contrast; a is a simplified parameter derived from the threshold inhibitor concentration m_{th} (the inflection point of the curve):

$$a = \frac{n+1}{n-1}(1 - m_{th})^n$$

In our model, $R_{min} = 0.003nm/s$, $R_{max} = 37.2nm/s$, $m_{th} = 0.686$, $n = 25.143$.

B Dataset Analysis

This appendix provides additional statistics for the FabBench datasets. Rather than only reporting sample counts, we analyze whether the datasets provide sufficient coverage and difficulty for metrology-aware evaluation. We organize the analysis around four questions: (1) what geometry and CD ranges are covered by the benchmark, (2) how CD and EPE distributions differ across datasets, (3) whether difficulty varies across pattern instances, and (4) how process conditions affect in-spec behavior and defect formation. All histogram-based panels use ratio-normalized vertical axes to enable comparison across datasets with different sample counts.

Geometry and CD coverage. We first examine whether the datasets cover diverse target dimensions. Figure 7 compares the target CD distributions for `TestPattern`, `OOD Process&Source`, and `MetalSet`. The top, bottom, and merged CD definitions provide complementary views of the intended geometry. A narrow target-CD distribution indicates a more concentrated design regime, while a broader distribution indicates more diverse linewidth and spacing

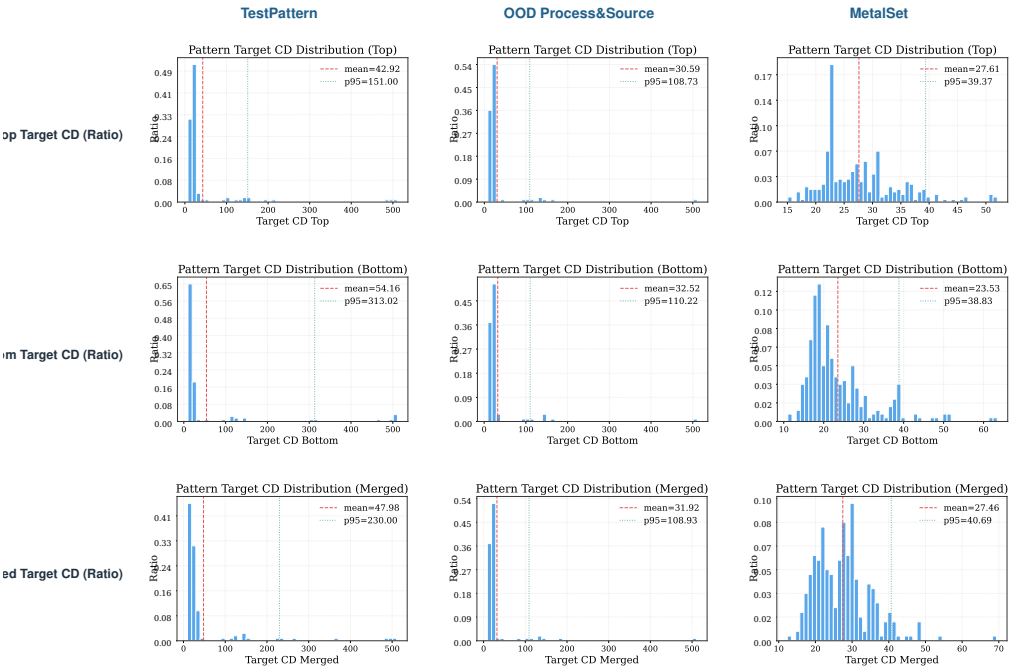


Figure 7: **Target CD distributions across datasets.** Ratio-normalized distributions of top, bottom, and merged target CDs for `TestPattern`, `OOD Process`, `OOD Source`, and `MetalSet`. These plots summarize the intended geometry coverage of each dataset.

conditions. This analysis verifies that the benchmark contains both controlled stress patterns and more regular design-layout statistics.

Figure 8 further reports the global CD distributions after process simulation and SEM rendering. Compared with target CD statistics, global CD reflects the effective printed response under process variation. The difference between target-CD and global-CD distributions helps reveal how lithography process effects reshape the intended mask geometry into printed SEM-domain observations.

EPE statistics and tail risk. We next analyze edge-placement behavior. Figure 9 reports core EPE statistics, which describe the typical edge-placement error distribution across the datasets. These statistics are useful for understanding the normal operating regime of each dataset and whether most samples remain within a moderate edge-displacement range.

However, average EPE alone can hide rare but important failures. Figure 10 therefore focuses on EPE tail behavior. Tail statistics are particularly relevant for lithography because a small number of severe edge-placement errors can dominate process-window risk. The ratio-normalized presentation makes it possible to compare the relative occurrence of high-risk samples across datasets despite different dataset sizes.

Pattern-level variability and intrinsic difficulty. Figure 11 analyzes variability at the pattern level. This is important because a dataset can have reasonable aggregate statistics while still containing a mixture of easy and process-sensitive patterns. The CD and EPE variability distributions indicate whether most patterns behave similarly or whether a small subset of patterns dominates metrology risk. This analysis also supports the interpretation of mask-OOD results: a held-out layout split is not necessarily harder in an absolute sense, because intrinsic pattern difficulty depends on layout regularity, local density, and critical-feature composition. In particular, our appendix analysis suggests that `MetalSet` is more stable under process variation than `TestPattern`, so relatively strong Mask-OOD performance should be interpreted as transfer to a held-out but comparatively easier mask distribution, rather than as evidence that the held-out split is intrinsically harder.

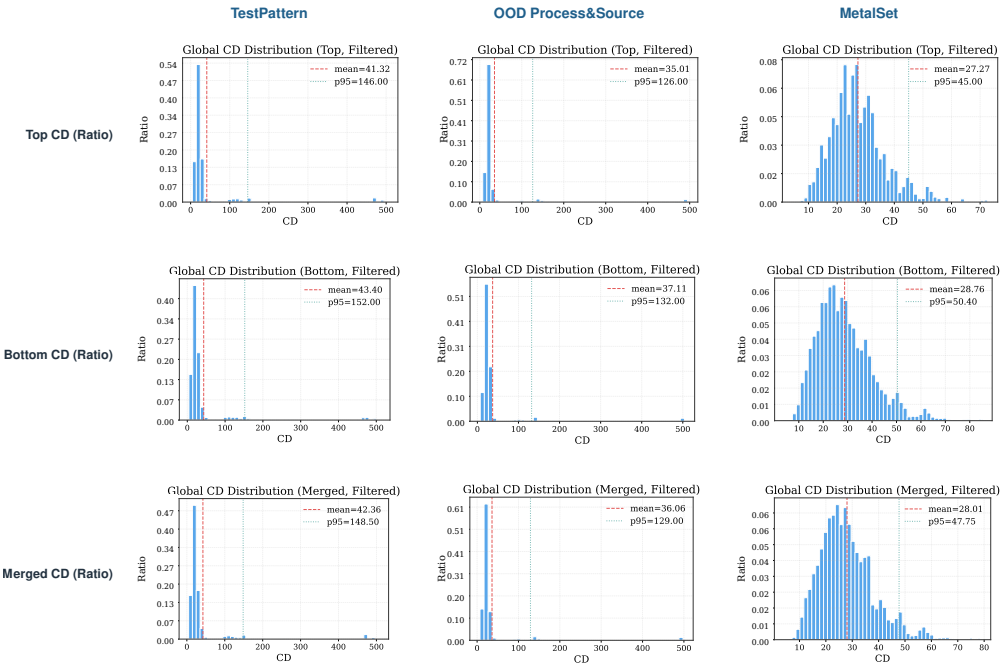


Figure 8: **Global CD distributions across datasets.** Ratio-normalized global CD distributions after process simulation and SEM rendering. These plots show how the intended geometry is transformed by process-dependent printing behavior.

In-spec behavior over the process window. To understand how samples respond to process perturbations, Figure 12 reports in-spec behavior across focus/exposure conditions. This analysis connects dataset statistics to process-window evaluation: patterns that remain in-spec across a wide process range are relatively stable, whereas patterns that fail under small perturbations are more process-sensitive. The top/bottom contour comparison is also useful for SEM-domain evaluation because the two contours can respond differently due to the white-edge effect and profile-sidewall geometry.

Defect composition. Figure 13 summarizes defect composition across datasets. FabBench includes multiple topology-level failure modes, including bridge, pinch, scumming, missing, and washout. The relative occurrence of these defects provides a complementary view of benchmark difficulty beyond continuous CD/EPE statistics. This is important because some models may preserve average geometry while failing to reproduce discrete defect states or defect types.

Process-condition sensitivity. Finally, Figure 14 summarizes process-condition sensitivity across datasets. We retain CD bias and structure-defect heatmaps because they provide the most interpretable view of process-window deformation. CD bias shows continuous metrology drift, while structure-defect maps show where process perturbations induce discrete topology failures. Together, these plots support the main benchmark design: process-OOD should be evaluated separately because changing process conditions can alter both continuous metrology quantities and discrete defect behavior.

Summary. Overall, these statistics show that FabBench provides more than a collection of SEM-like images. The datasets differ in intended CD coverage, printed CD response, EPE distribution, pattern-level variability, process-window stability, and defect composition. In particular, the stability gap between `TestPattern` and `MetalSet` helps contextualize the Mask-OOD results: transfer to held-out masks in FabBench should be interpreted together with the relative difficulty of the target

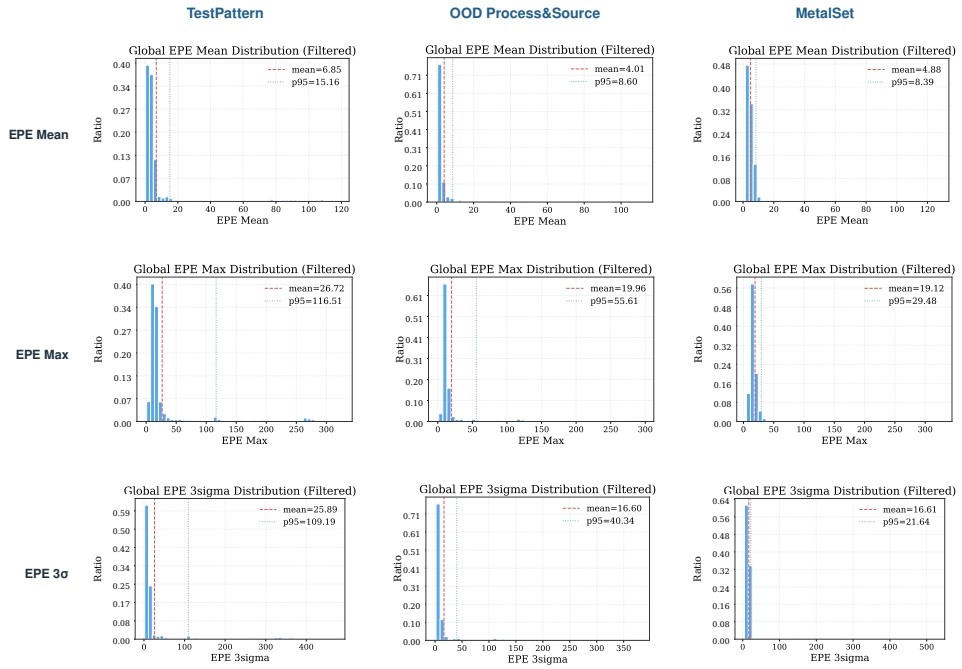


Figure 9: **Core EPE statistics across datasets.** The panels summarize typical edge-placement behavior for each dataset, providing a view of the common metrology-error regime before focusing on extreme tails.

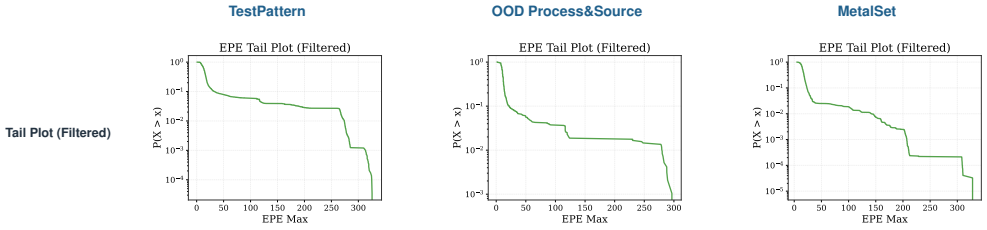


Figure 10: **EPE tail-risk analysis.** Ratio-normalized tail statistics highlight rare but practically important edge-placement failures. These tails are important because severe local EPE errors can dominate process-window risk even when average error remains moderate.

distribution. This supports the use of FabBench as a controlled benchmark for evaluating not only visual SEM prediction, but also metrology behavior, defect sensitivity, and process robustness.

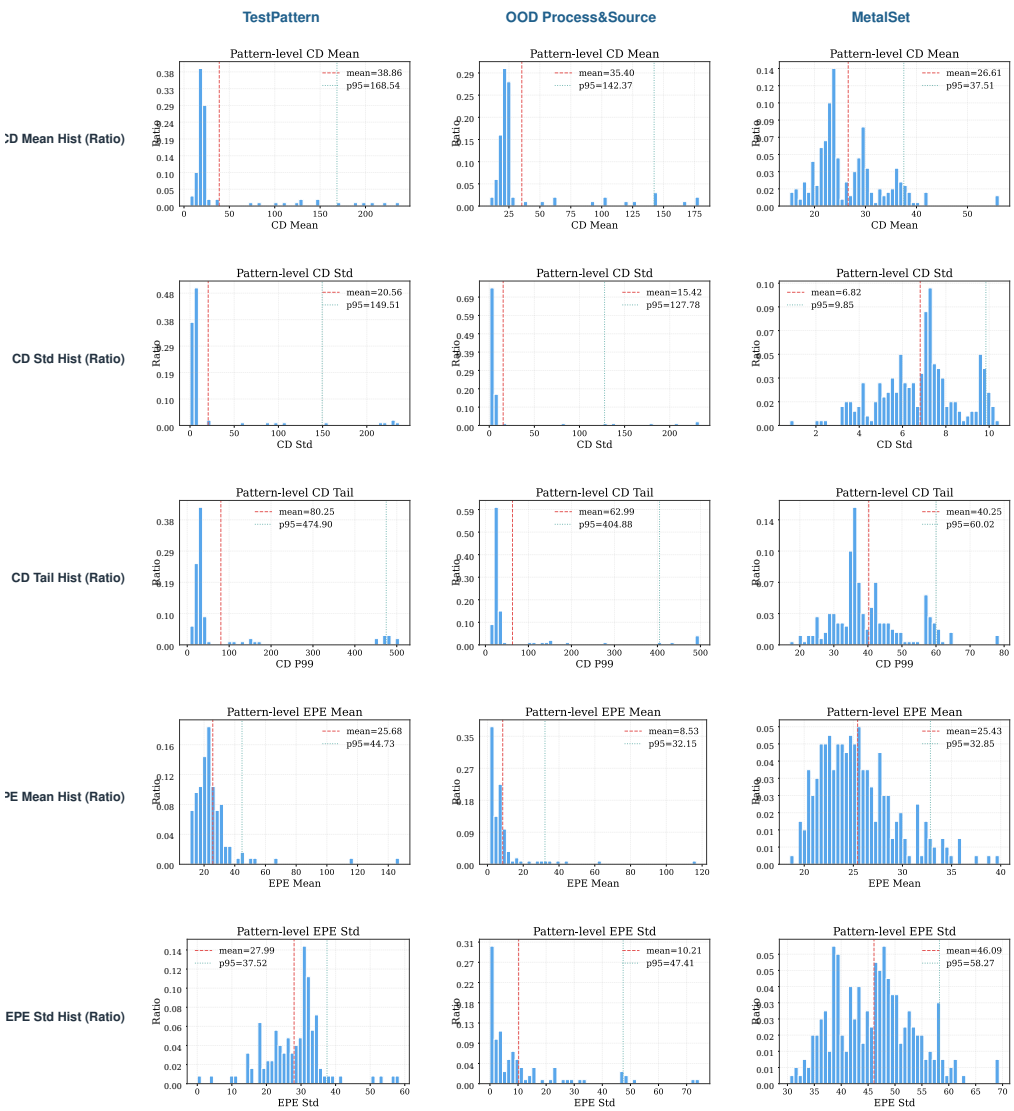


Figure 11: **Pattern-level CD and EPE variability.** Histogram rows use ratio-normalized vertical axes, while boxplots summarize inter-pattern dispersion. These plots characterize intrinsic pattern difficulty and show whether metrology risk is uniformly distributed or concentrated in process-sensitive patterns.

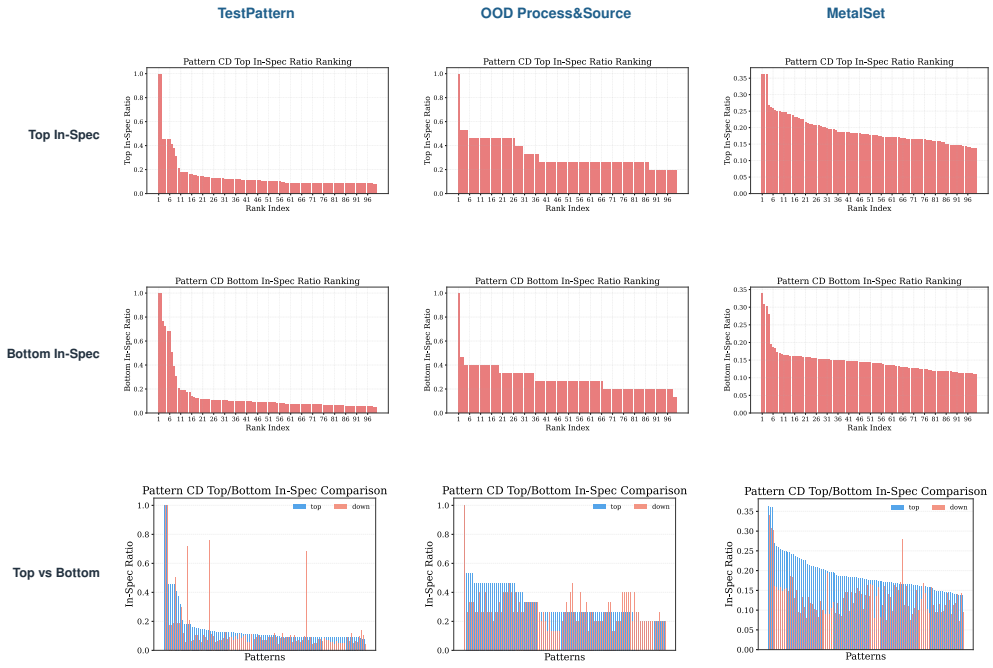


Figure 12: **In-spec behavior across process conditions.** The panels summarize how top and bottom contour measurements remain in-spec or fail across focus/exposure conditions. This reveals process-window stability and supports the use of dual-contour metrology in FabBench.

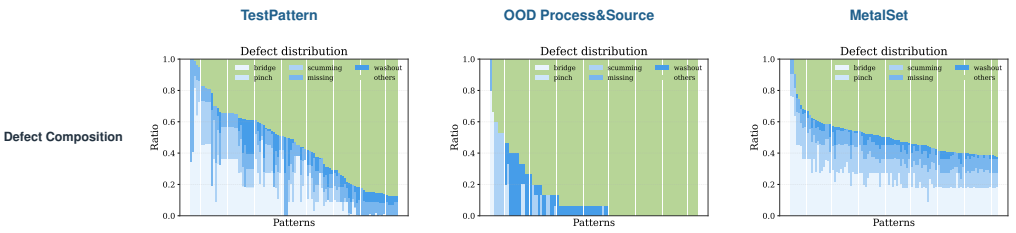


Figure 13: **Defect composition across datasets.** The panels report the relative occurrence of bridge, pinch, scumming, missing, and washout defects. These statistics characterize topology-level failure behavior beyond continuous CD and EPE distributions.

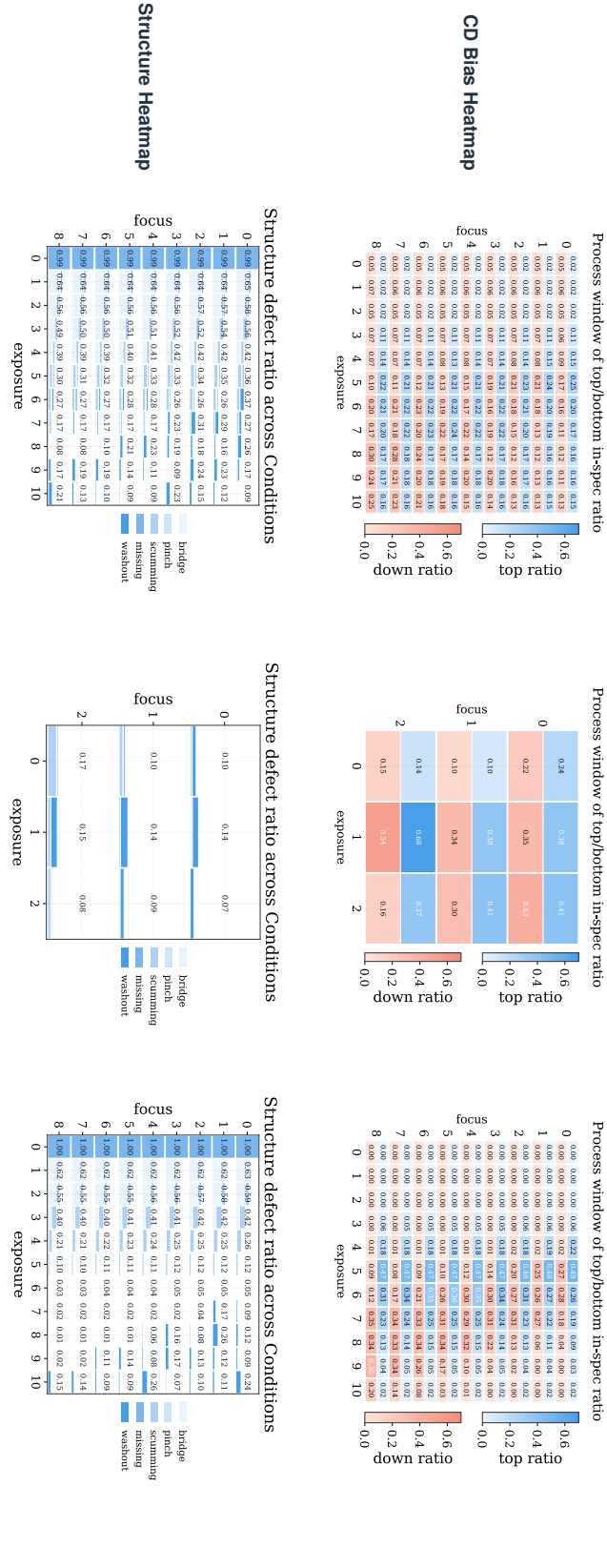


Figure 14: **Process-condition sensitivity across datasets.** CD bias and structure-defect heatmaps highlight how each dataset responds to focus/exposure perturbations. CD bias captures continuous metrology drift, while structure-defect maps reveal process-induced topology failures.

C SEM EMULATOR DETAILS AND VALIDATION

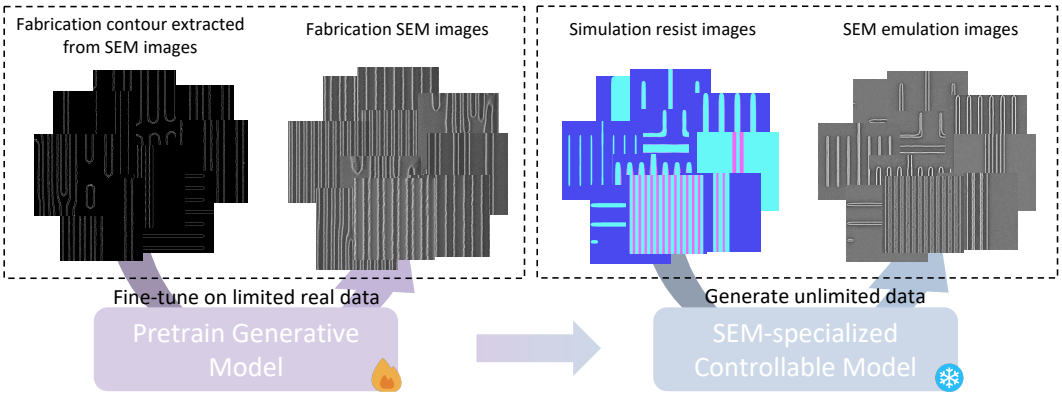


Figure 15: **SEM emulation framework.** The emulator learns a mapping from geometric contour/profile representations to SEM-style observations, transferring physics-simulated resist geometry into the SEM appearance domain.

Table 5: **Overview of the real SEM dataset used for SEM emulator training and validation.**

Item	Specification
Mask Patterns	200 unique layouts
Process Step	After-develop inspection images
Focus Range	Nominal focus $\pm\Delta F$ (9 steps)
Dose Range	Nominal dose $\pm\Delta D$ (11 steps)
Source Family	1 illumination family
Total Images	20k real CD-SEM clips
Train/Test Ratio	9:1
Technology Node	14–28nm
Metrology Tool	KLA CD-SEM G* series

Rigorous lithography simulation provides process-dependent resist profiles, while fabrication-side inspection is typically performed using CD-SEM images. To connect these two domains, we develop an SEM emulator that renders simulated resist geometry into SEM-like metrology observations. The emulator is implemented using ControlNet with a Stable Diffusion 1.5 backbone. Its role is not to replace calibrated wafer SEM measurement, but to introduce real-SEM-informed appearance characteristics, such as edge brightening, local contrast variation, and stochastic texture, into the benchmark targets.

This emulator enables scalable generation of spatially aligned SEM-like samples from physics-computed resist profiles. Since the underlying simulation provides precise geometry and process metadata, the generated targets can be paired with contour, defect, and gauge-based metrology annotations. This design allows FabBench to construct a controlled SEM-domain benchmark while avoiding the practical limitations of fully public fab-scale SEM datasets, which are often proprietary, sparse, irregularly sampled, and difficult to annotate exhaustively.

Training. A key challenge in learning SEM emulation is preventing the model from memorizing specific focus–dose conditions or layout instances. To encourage the emulator to learn a geometry-to-appearance mapping, we construct training pairs from real SEM data by extracting contours from SEM images and pairing them with the corresponding SEM observations. In other words, the model is trained on $(\text{contour}, \text{SEM})$ pairs rather than directly on process-condition labels. This decouples SEM appearance learning from the specific process settings used to acquire each image and encourages the emulator to model how local geometric morphology is rendered into SEM texture.

Table 6: **Real-data validation of the SEM emulator.** The emulator is evaluated on held-out real SEM samples using contour, metrology, and distributional metrics.

Method	Dice↑	B-IoU↑	EPE(nm)↓	CD(nm)↓	Status F1↑	Type F1↑	FID↓	KID↓
SEM Emulator	98.84	92.17	1.24	2.12	96.21	93.10	31.36	0.0086 ± 0.0024

As shown in Fig. 15, the extracted contour representation serves as the spatial condition for ControlNet, while the diffusion backbone synthesizes SEM-style intensity patterns. The real SEM dataset used for training and validation is summarized in Table 5.

Inference During inference, the emulator takes contours extracted from rigorous 3D resist-profile simulations as input. To faithfully replicate SEM image artifacts, we identify regions with high sidewall slopes and apply adaptive noise injection and brightness enhancement. We categorize the stochastic artifacts into four distinct modes: $\mathcal{S} = \{\text{Fill-in}, \text{Random noise}, \text{Random line}, \text{None}\}$. The selection of the specific noise mode $s \in \mathcal{S}$ is governed by the sidewall integrity of the resist profile, quantified by the effective projection ratio (ρ):

$$\rho = \frac{\max\left(\mathcal{A}(C_{\text{bottom}} \cup C_{\text{top}}) - \mathcal{A}(C_{\text{top}}), 0\right)}{\mathcal{A}(M_{\text{target}})} \tag{7}$$

where $\mathcal{A}(\cdot)$ denotes the area operator, and $C_{\text{bottom}}, C_{\text{top}}$ represent the resist contours at the substrate and top surface, respectively. The term $\mathcal{A}(C_{\text{bottom}} \cup C_{\text{top}}) - \mathcal{A}(C_{\text{top}})$ indicates the exposed sidewall projection area visible from a top-down perspective:

- Normal Taper ($C_{\text{top}} \subset C_{\text{bottom}}$): The formula yields $\mathcal{A}(C_{\text{bottom}}) - \mathcal{A}(C_{\text{top}})$, representing the visible sidewall region that contributes white-edge SEM signal.
- Near-zero-projection regimes: when the visible sidewall projection becomes negligible, the resulting ratio approaches $\rho \approx 0$. This can arise in more than one physical situation, including near-vertical profiles ($C_{\text{bottom}} \approx C_{\text{top}}$) and occluded profiles such as inverse taper, overhang, or collapse ($C_{\text{bottom}} \subseteq C_{\text{top}}$).

Accordingly, ρ is used here as an observation-oriented heuristic for the amount of visible sidewall signal in top-down SEM, rather than as a unique classifier of resist-profile state. In particular, $\rho \approx 0$ indicates that the SEM image should contain little or no visible white-edge evidence, even though the underlying physical profile may not be uniquely identifiable from top-bottom SEM alone.

We map ρ to the noise injection modes based on the correlation between visible sidewall projection and electron signal yield:

- High ρ (Significant tapering): Assigned the **Fill-in** mode. A large difference between bottom and top contours implies a degraded lithographic profile (shallow sidewall angle), resulting in the strongest visual signal (wide white bands).
- Intermediate ρ: Sequentially assigned to **Random Noise** and **Random Line**. This represents standard profiles where moderate sidewall slopes generate typical edge roughness and signal scattering.
- Low ρ / negligible visible sidewall projection: Assigned minimal-signal modes. In top-bottom SEM, this regime includes both near-vertical profiles and occluded profiles, since both can produce little visible white-edge evidence.
- At the weakest-signal end of this regime, we use the **None** mode. This choice reflects that additional injected artifacts would otherwise overstate sidewall visibility in cases where the top-bottom SEM image provides little or no observable white-edge signal.

Ablation study of Adaptive Noise Injection Strategy The ablation study of the adaptive noise injection strategy is shown in Figure 16, where the None mode can reduce fidelity when resist profiles exhibit edge roughness.

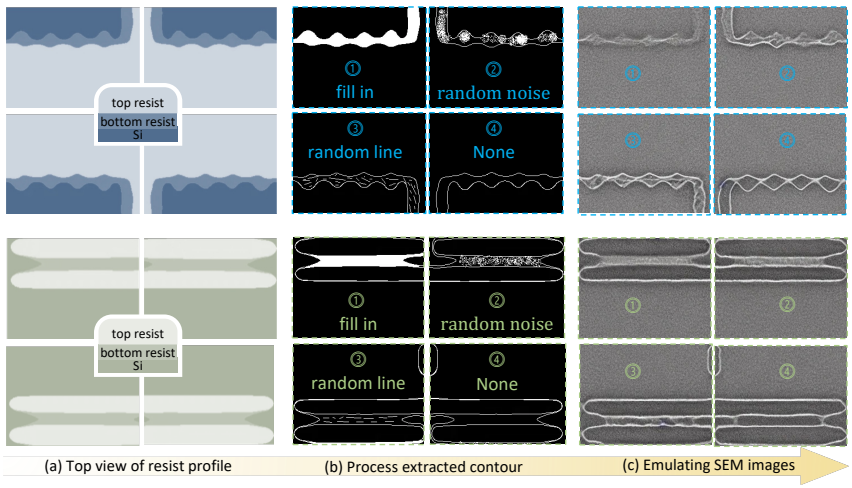


Figure 16: **Ablation study of adaptive noise injection strategy.** The injection mode is dynamically determined by the pattern integrity ratio (ρ) to mimic realistic electron scattering artifacts.

SEM Emulator Fidelity on Held-out Real SEM Because FabBench benchmark targets are generated by a learned SEM emulator, we first evaluate the emulator on held-out real SEM samples. As shown in Table 6, the emulator achieves strong contour and metrology agreement with real SEM observations, with Dice of 98.84, B-IoU of 92.17, EPE of 1.24, and CD error of 2.12. We also measure the distributional gap between emulator outputs and real SEM images, obtaining an FID of 31.36 and a KID of 0.0086 ± 0.0024.

These results indicate that the SEM emulator preserves the major observation-domain structures needed for benchmark construction, including contour geometry, edge placement, and SEM-style appearance. Nevertheless, the emulator is not treated as a replacement for wafer-measured SEM ground truth; rather, it provides a controlled and real-SEM-informed rendering module for scalable benchmark generation.

Table 7: **SEM emulator adaptation to a neighboring inspection process.** Full fine-tuning adapts efficiently with limited target-domain data, while LoRA is weaker in this setting.

Method	Data	Eff. Epochs	Best KID↓	MAE↓
Full FT	25%	**3.75**	0.0038	0.0666
Full FT	100%	6.00	**0.0015**	**0.0652**
LoRA	50%	23.50	0.0076	0.1002
LoRA	100%	24.00	0.0079	0.1011

Cross-domain Adaptation of the SEM Emulator We further evaluate whether the real-SEM grounding of the emulator can be adapted to a neighboring inspection process. This experiment tests whether the emulator must be fixed to a single imaging domain, or whether it can be recalibrated when inspection conditions change. As reported in Table 7, full fine-tuning with only 25% of the target-domain data reaches a KID of 0.0038 and an MAE of 0.0666 after 3.75 effective full epochs. This is close to the 100%-data full fine-tuning setting, which achieves the best KID of 0.0015 and MAE of 0.0652 after 6.00 effective epochs.

In contrast, LoRA adaptation is weaker in this setting: both 50% and 100% LoRA runs require over 23 effective epochs and remain worse in both KID and MAE. These results suggest that the SEM emulator can be rapidly recalibrated with modest additional real SEM data when the inspection regime changes, although the most effective adaptation strategy may depend on the target domain.

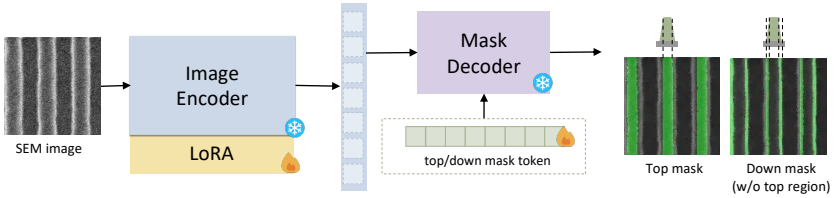


Figure 17: **Architecture of the Proposed SAM-based Evaluator.** We employ Low-Rank Adaptation (LoRA) to adapt the ViT encoder to the SEM visual modality. In the mask decoder, we introduce task-specific learnable tokens (Top/bottom) to resolve semantic ambiguity between the resist pattern and the substrate background.

D FINE-TUNE SAM FOR AUTO-EVALUATION

Encoder adaptation via LoRA The Segment Anything Model (SAM) employs a Vision Transformer (ViT-H/L/B) encoder, which encapsulates rich geometric priors from large-scale pre-training. However, a significant domain gap exists between natural images and the high-frequency, noisy modality of SEM images. To adapt this powerful backbone to the lithography domain while preserving its generalized segmentation capabilities, we employ Low-Rank Adaptation (LoRA). Specifically, within each Transformer block, the input features are processed via Multi-Head Self-Attention (MHSA). Instead of fine-tuning the massive pre-trained weights—which risks catastrophic forgetting—we inject trainable low-rank decomposition matrices parallel to the frozen query ($\mathbf{W}_q$) and value ($\mathbf{W}_v$) projection layers. Mathematically, for a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times d}$, the update is constrained as:

$$W = W_0 + \Delta W = W_0 + BA \tag{8}$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times d}$ are low-rank matrices with rank $r \ll d$. This strategy enables the model to efficiently capture unique SEM textures and noise distributions with minimal computational overhead.

Semantic Disambiguation via Learnable Tokens While SAM excels at class-agnostic segmentation, it inherently lacks semantic understanding, tending to segment any prominent object regardless of its relevance. In SEM imaging of semiconductor layers, this poses a critical issue: the model often struggles to distinguish between the target resist patterns (top layer) and the underlying substrate or background noise (bottom layer), as both appear as distinct regions. To decouple these overlapping semantics, we introduce two specific learnable semantic tokens into the mask decoder, designated as the Top-Mask Token and Bottom-Mask Token. Unlike standard geometric prompts (points/boxes), these tokens serve as global semantic queries. During training, we supervise the Top-Mask Token using the ground-truth top/bottom resist images derived from rigorous simulation, enforcing the model to align this token specifically with the resist features. During inference, injecting the trained Top-Mask Token automatically prompts the decoder to suppress background artifacts and selectively segment the target resist geometry.

Table 8: **Held-out validation accuracy of the fine-tuned SAM extractor.** The extractor is evaluated against expert-calibrated contour and defect annotations.

Method	Dice↑	B-IoU↑	EPE(nm)↓	CD(nm)↓	Status F1↑	Type F1↑
SAM Extractor	99.67	96.13	0.20	0.45	98.27	96.70

Contour Extraction Accuracy FabBench uses a fine-tuned SAM module to extract top and bottom contours from SEM-domain observations. Since contour-derived geometric metrics, CD, EPE, and rule-based defect labels all depend on this extraction step, we evaluate the SAM extractor separately on a held-out expert-calibrated validation set. This validation set is manually checked before evaluation. Specifically, contour candidates and initial defect tags are first derived from the rigorous lithography simulator and topology rules, and lithography-domain experts then inspect each SEM

image with the mask overlay to correct local boundary errors, missing/extra components, top/bottom semantic ambiguities, and incorrect defect assignments. The finalized expert-calibrated contours and defect tags are used as the reference annotations for evaluating SAM extraction accuracy, CD/EPE measurement consistency, and defect-label consistency under the automated extraction pipeline. As shown in Table 8, the fine-tuned SAM extractor achieves an average Boundary IoU of 96.13%, EPE of 0.20, and CD error of 0.45, defect status F1 of 98.27% and type F1 of 96.70%. These results indicate that the automated contour extraction protocol is reliable for consistent evaluation of both benchmark targets and model predictions. The high contour agreement supports using the SAM-based extractor as a consistent automated metric backbone for comparing benchmark targets and model predictions.

E Evaluation metric details

E.1 Metrics definition

FabBench evaluates SEM-domain predictions using a coarse-to-fine metric suite: image-level fidelity, contour geometry, local metrology, defect sensitivity, and efficiency. All contour-, metrology-, and defect-related metrics are computed after applying the same fine-tuned SAM-based contour extraction pipeline to both the predicted SEM image and the benchmark target. This shared evaluator ensures that metric computation is consistent across methods and does not depend on model-specific post-processing.

Image-level fidelity. We report PSNR, LPIPS, and FID to assess SEM-domain appearance fidelity. PSNR measures pixel-level reconstruction quality, LPIPS measures perceptual distance in a learned feature space, and FID measures dataset-level distributional similarity between predicted and benchmark SEM-domain images. These metrics capture overall image appearance, but they do not directly measure whether the prediction preserves fabrication-relevant geometry, local metrology, or defect behavior.

SAM-based contour extraction. To evaluate geometry and metrology from SEM-domain images, FabBench extracts both top and bottom contours using a fine-tuned Segment Anything Model (SAM). This step is necessary because SEM images contain observation-domain effects such as edge brightening and texture, so visible SEM boundaries do not perfectly coincide with ideal simulator contours. The SAM extractor therefore serves as a unified evaluator that maps both benchmark targets and model predictions into comparable top/bottom contour representations. Fine-tuning and validation details are provided in Appendix D.

To assess the reliability of this automated extraction step, we evaluate the SAM extractor on a held-out contour test set. The test annotations are initialized from rigorous-simulation contours and manually refined to correct local boundary errors and top/bottom contour ambiguities. The extractor achieves high contour and metrology consistency, including a mean CD error of $0.45\,\mathrm{nm}$; full validation results are provided in Appendix D.

Contour geometry. Let B^c and $\hat{B}^c$ denote the benchmark and predicted binary masks for contour type $c \in \{\mathrm{top}, \mathrm{bottom}\}$. We compute Mask IoU and Dice score as

$$\mathrm{IoU}^c = \frac{|B^c \cap \hat{B}^c|}{|B^c \cup \hat{B}^c|}, \tag{9}$$

$$\mathrm{Dice}^c = \frac{2|B^c \cap \hat{B}^c|}{|B^c| + |\hat{B}^c|}. \tag{10}$$

These region-based metrics measure coarse contour agreement between the prediction and benchmark target.

We further report Boundary IoU to emphasize boundary-level geometry. Unlike standard Mask IoU, which measures overlap over the full printed region, Boundary IoU computes overlap only within narrow bands around the contour:

$$\mathrm{BoundaryIoU} = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|}, \tag{11}$$

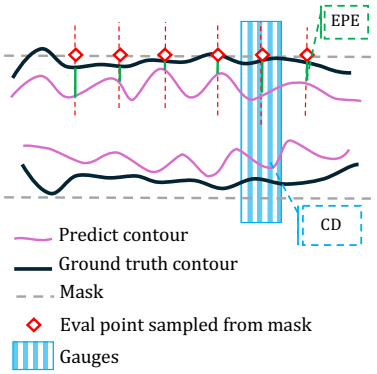


Figure 18: **Schematic of gauge-based CD and EPE measurements.** CD is measured at predefined metrology gauges from the extracted top/bottom contours, while EPE quantifies local edge displacement between the predicted and benchmark target contours along sampled edge locations.

where G and P denote the benchmark and predicted binary masks, respectively, and G_d and P_d denote the sets of pixels within Euclidean distance d from the corresponding contours. Boundary IoU is therefore more sensitive to boundary displacement, corner rounding, and local shape deformation than full-region overlap.

For all contour metrics, we compute the score separately on top and bottom contours and report their average unless otherwise specified.

Gauge-based CD error. Critical Dimension (CD) measures local linewidth at predefined metrology gauges. FabBench adopts mask-anchored gauges generated by a hybrid protocol that combines rule-based automated generation with manual verification. Gauges are designed to capture process-sensitive regions while avoiding ambiguous locations such as 2D corners or severe topological breakdown regions where 1D linewidth measurement is ill-posed. Each gauge is stored with its spatial coordinate, measurement direction, and nominal CD in a standardized JSON schema.

During evaluation, each mask-defined gauge coordinate is deterministically mapped to the SEM spatial domain. At a valid gauge j, CD is computed as the intersection distance between the two corresponding contour edges along the gauge direction. Because SEM observations exhibit white-edge effects and can reflect sidewall asymmetry, CD is measured on both top and bottom contours. The dual-contour CD error is defined as

$$\mathrm{CD}_{\mathrm{err}} = \frac{1}{2N} \sum_{j=1}^{N} \left(\left| \widehat{\mathrm{CD}}_j^{\mathrm{top}} - \mathrm{CD}_j^{\mathrm{top}} \right| + \left| \widehat{\mathrm{CD}}_j^{\mathrm{bottom}} - \mathrm{CD}_j^{\mathrm{bottom}} \right| \right), \tag{12}$$

where N is the number of valid gauge measurements. A lower CD error indicates better preservation of local linewidth behavior.

Edge Placement Error (EPE). EPE measures local edge displacement between the predicted and benchmark target contours. For each sampled edge location or gauge-associated sampling point j, we compute the absolute displacement between the predicted contour and the corresponding benchmark contour along the predefined measurement direction. Similar to CD, EPE is computed on both top and bottom contours. Let Δ_j^{top} and $\Delta_j^{\mathrm{bottom}}$ denote the absolute contour-to-contour displacements for the top and bottom contours. The dual-contour EPE is defined as

$$\mathrm{EPE}_{\mathrm{err}} = \frac{1}{2N} \sum_{j=1}^{N} \left(\Delta_j^{\mathrm{top}} + \Delta_j^{\mathrm{bottom}} \right). \tag{13}$$

A lower EPE indicates more accurate local edge placement. Together, CD and EPE provide complementary local metrology views: EPE measures edge displacement, while CD measures the spacing between paired edges.

Defect-level sensitivity. FabBench evaluates whether predictions reproduce process-induced topological failures. Defect labels are derived from the extracted top/bottom contours and the corresponding mask using deterministic topology rules. All masks and contours are converted to binary masks, and connected components are computed with 8-connectivity. No additional morphological opening or closing is applied during defect assignment.

Let $\mathcal{C}_M$, $\mathcal{C}_T$, and $\mathcal{C}_B$ denote the connected components of the mask, top contour, and bottom contour masks, respectively. A *bridge* is detected if one top component overlaps at least two distinct mask components. To avoid classifying global collapse as bridging, this rule is disabled when a top component covers at least 90% of the image area. A *pinch* is detected if one mask component overlaps at least two distinct top components. A *scumming* defect is detected from the residual bottom-only region $B \setminus T$: if one connected component of $B \setminus T$ overlaps at least two distinct mask components, it is labeled as scumming. A *missing* defect is detected in either of two cases: (i) a top component covers at least 90% of the image area, or (ii) the top and bottom masks nearly collapse into the same image-scale region, defined by

$$\frac{|(T \cup B) \setminus (T \cap B)|}{|T \cup B|} < 0.02 \quad \text{and} \quad \frac{|T \cup B|}{|\Omega|} \geq 0.90, \tag{14}$$

where Ω is the image domain. A *washout* defect is detected if both top and bottom masks disappear inside a non-empty mask, or if a bottom component overlaps the target mask but has zero overlap with the top mask.

These rules produce five non-exclusive defect categories: bridge, pinch, scumming, missing, and washout. When several rules fire for one sample, FabBench retains every activated category as a multi-hot prediction rather than collapsing the result to a single priority label.

We report two F1-based metrics. Defect-status F1 treats defect evaluation as binary classification between in-spec and defective samples:

$$\mathrm{F1}_{\mathrm{status}} = \frac{2 \cdot \mathrm{Precision}_{\mathrm{status}} \cdot \mathrm{Recall}_{\mathrm{status}}}{\mathrm{Precision}_{\mathrm{status}} + \mathrm{Recall}_{\mathrm{status}}}. \tag{15}$$

Defect-type F1 evaluates whether the predicted label set matches the multi-label benchmark annotation. For each class c, we compute

$$\mathrm{F1}_c = \frac{2\mathrm{TP}_c}{2\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}. \tag{16}$$

The macro and micro variants are

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{5}\sum_{c=1}^{5} \mathrm{F1}_c, \qquad \mathrm{F1}_{\mathrm{micro}} = \frac{2\sum_c \mathrm{TP}_c}{2\sum_c \mathrm{TP}_c + \sum_c \mathrm{FP}_c + \sum_c \mathrm{FN}_c}. \tag{17}$$

Macro-F1 weights the five defect categories equally, whereas micro-F1 pools all sample–class decisions. Classes with zero ground-truth support are retained in the macro average, with undefined per-class terms set to zero. Main Tables 2 and 3 report Type F1 as macro/micro. Defect-status F1 measures whether a model detects any structural failure, while multi-label defect-type F1 measures which failure modes it reproduces.

Turnaround Time (TAT). We report turnaround time (TAT) to evaluate inference efficiency for large-scale process-window exploration. TAT is defined as the average wall-clock inference latency required to synthesize one SEM-domain sample per GPU, computed as the total elapsed generation time divided by the number of generated samples. All TAT values are measured on a single NVIDIA A800 GPU with batch size 10 at the benchmark resolution used for all reported models. For diffusion-based baselines, the timing includes the full 50-step denoising trajectory used for benchmark evaluation. TAT includes model conditioning and forward SEM synthesis, but excludes data loading, offline preprocessing, SAM-based contour extraction, and metric computation. A lower TAT indicates higher inference throughput.

E.2 Metrics analysis

To analyze the behavior of the proposed metric suite, we apply controlled distortions to representative test patterns, including dilation, erosion, shift, noise, and holes. As shown in Fig. 19, region-overlap metrics such as Mask IoU and Dice mainly capture coarse shape agreement, whereas Boundary IoU is more sensitive to boundary-localized distortions such as dilation, erosion, and shifts. This

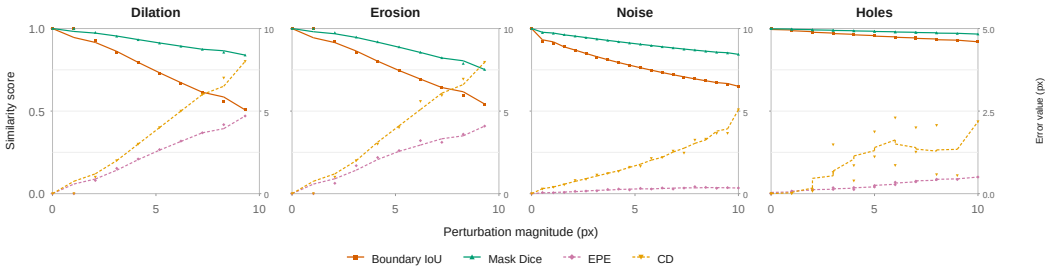


Figure 19: Metric sensitivity under controlled distortions. The analysis illustrates why visually plausible or high-overlap predictions can still fail local metrology or defect-sensitive evaluation.

makes Boundary IoU useful for detecting contour-level shape changes that may be weakly reflected in full-region overlap.

CD and EPE provide complementary local metrology views. EPE measures edge displacement, while CD measures the distance between paired contour edges along predefined gauge directions. Therefore, a prediction can have accurate edge placement in some regions while still producing incorrect local linewidth, or vice versa. This distinction explains why high contour overlap or low EPE does not necessarily imply accurate CD.

Defect metrics further probe discrete topology changes that may not be captured by continuous geometric or metrology errors alone. For example, localized holes or connectivity changes can have limited effect on region-level overlap while changing the defect status or defect type. Together, the metric suite evaluates SEM-domain predictions from image appearance to contour geometry, local metrology, and topology-level failure behavior, supporting the fabrication-aligned analysis used in FabBench.

F External wafer-SEM CD check

We conduct an external sanity check using real measured SEM CD values that are not part of the released FabBench benchmark targets. The goal is to examine whether FabBench's CD metric shows broadly consistent behavior with an independently measured fabrication-side CD signal.

The validation subset contains six focus-wise dose curves. These curves are selected from a held-out wafer-SEM subset that is disjoint from the real SEM images used to train, validate, or adapt the SEM emulator. The corresponding masks and focus-dose measurements are also not used for FabBench target generation, baseline model training, model selection, or hyperparameter tuning. This separation is intended to prevent leakage from SEM-emulator training data into the external measured-SEM sanity check.

For each model, we compare the predicted CD curve with the measured wafer SEM CD curve at matched focus-dose conditions. We report two measured-SEM quantities: (i) CD Bias, the mean absolute prediction–measurement offset at the manually measured gauge locations, and (ii) Trend Corr., the mean Pearson correlation across valid focus-wise dose curves.

To keep this external check independent from the benchmark evaluator, the measured-SEM CD values are manually measured at predefined gauge locations from the wafer SEM clips. They are not extracted by the FabBench SAM-based contour evaluator and are not obtained through any model-specific post-processing. This avoids coupling the measured-SEM sanity check to the same automated evaluator used in the main benchmark and makes the comparison a separate fabrication-side metrology check.

To compare FabBench with this measured-SEM signal, we use each model's Mask-OOD CD error, which is the closest FabBench split to this held-out validation. The matched model mapping is CNN, Trans, Pix2Pix, SD-Adapter, RFNO, and MFNO. Qualitatively, models with lower FabBench CD error tend to better follow the measured SEM CD trends, while models with larger FabBench CD error show larger measured-SEM CD discrepancy or weaker trend tracking. This observation

Table 9: **External wafer-SEM CD check against FabBench CD error.** The "FabBench CD" column reports each model's CD error on the FabBench Mask-OOD split from the main benchmark table; the remaining columns report measured-SEM CD quantities computed on the held-out wafer SEM subset. The measured-SEM CD values are manually measured at fixed gauge locations and are not produced by the FabBench SAM evaluator. CD Bias is the mean absolute prediction–measurement offset, and Trend Corr. is the mean Pearson correlation across valid focus-wise dose curves. These per-model summaries are used only for a descriptive CD-focused sanity check.

Model	FabBench CD ↓	CD Bias ↓	Trend Corr. ↑
Pix2Pix	11.17	2.685	0.969
RFNO	21.33	7.804	0.246
CNN	24.22	12.396	−0.399
Trans	41.55	18.753	−0.756
MFNO	19.05	3.483	0.980
SD-Adapter	26.98	7.843	0.789

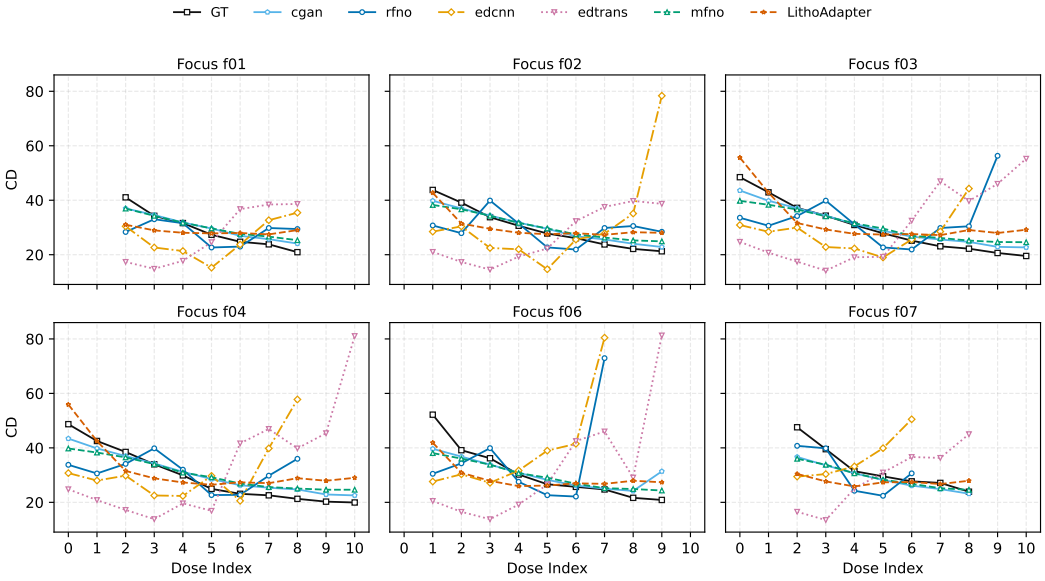


Figure 20: **External wafer-SEM CD trend check.** Measured wafer SEM CD curves and model-predicted CD curves are compared across valid focus-wise dose sweeps. The plot is used to inspect directional CD trend consistency only, not as benchmark ground truth or as full measured-SEM validation of EPE, IoU, or defect behavior.

provides a small external check that FabBench's CD metric captures a metrology-relevant signal, but it should not be interpreted as statistically conclusive validation.

This pilot check is intentionally CD-focused. CD is the most stable gauge-based measurement available in this measured-SEM subset. EPE requires denser boundary-aligned annotations and more reliable contour registration, while defect-type F1 requires substantially more defect-positive samples per class. We therefore do not report measured-SEM EPE or defect-type F1 on this small subset, and leave those validations to larger future measured-SEM releases.

The raw wafer SEM clips, process recipes, calibration parameters, registration metadata, and sensitive mask information are not released because of proprietary and manufacturing-sensitivity constraints. We report only anonymized focus/dose indices and aggregate CD statistics. This measured-SEM sanity check is used as external supporting evidence for FabBench's CD signal, not as benchmark ground truth or as production qualification evidence.

Table 10: Cross-dataset model-ranking consistency. Each entry gives the FabBench rank followed by the wafer-SEM rank.

Model	CD	EPE	B-IoU	FID	Status F1
CNN	$4/5$	$2.5/2$	$4/2$	$4/3$	$4/5$
Transformer	$6/6$	$5/5$	$6/6$	$6/6$	$6/6$
RFNO	$3/3$	$6/6$	$5/5$	$5/5$	$5/4$
MFNO	$2/2$	$4/4$	$3/4$	$3/4$	$3/3$
Pix2Pix	$1/1$	$2.5/3$	$2/3$	$2/2$	$2/2$
SD-Adapter	$5/4$	$1/1$	$1/1$	$1/1$	$1/1$
Spearman ρ	0.943	0.986	0.829	0.943	0.943

Table 11: Stress test on dense-mask layouts. EPE is reported in nanometers.

Model	EPE ↓	Dice ↑	B-IoU ↑
CNN	13.55	0.711	0.237
Transformer	15.65	0.719	0.139
RFNO	12.90	0.743	0.241
MFNO	13.11	0.753	0.247
Pix2Pix	**10.81**	**0.849**	**0.336**
SD-Adapter	11.35	0.765	0.288

G Additional experiments and analyses from the author response

This section records the additional quantitative results reported during the author-response period. These results supplement the original submission without replacing its benchmark protocol or its original tables. Unless stated otherwise, the evaluation definitions are identical to those in the main paper. The benchmark images use a fixed scale of $1\,\mathrm{nm/pixel}$; therefore, the numerical CD and EPE values are unchanged when the unit in the main tables is corrected from pixels to nanometers.

G.1 Evaluation on held-out wafer SEM

We evaluate the trained predictors directly against independently held-out wafer SEM images, so the SEM emulator is not used as the evaluation target. The main-paper Table 4 reports the resulting fabrication-relevant and appearance-oriented metrics. Pix2Pix gives the lowest CD error, whereas SD-Adapter gives the lowest EPE and FID and the highest B-IoU and defect-status F1. The boundary–metrology discrepancy observed on FabBench is also present on wafer SEM: for example, CNN attains the second-highest B-IoU but has a comparatively large CD error.

Table 10 compares model ranks on FabBench and wafer SEM. Spearman rank correlations range from 0.829 to 0.986 across the five metrics, indicating that the benchmark preserves broad relative differences among the evaluated model families despite the residual emulator-to-fabrication domain gap.

G.2 Dense-layout stress test

Table 11 extends the topology-transfer evaluation to severe dense layouts. The results amplify the gap between coarse overlap and local boundary fidelity: Transformer obtains a moderate Dice score of 0.719 but the lowest B-IoU and the largest EPE among the evaluated models.

G.3 Multi-label defect evaluation

Main Tables 2 and 3 report multi-label macro/micro Type F1 for every evaluated baseline. Macro-F1 is the unweighted mean of the five per-class F1 scores, while micro-F1 pools true positives, false positives, and false negatives across all classes. Classes with zero ground-truth support are retained

Table 12: Defect distribution across splits. Percentages are relative to the analyzed samples in each row.

Split	Samples	Bridge	Pinch	Scumming	Missing	Washout
ID	98,851	14,794 (14.97%)	192 (0.19%)	12,681 (12.83%)	10,191 (10.31%)	5,248 (5.31%)
Mask-OOD	445,499	54,110 (12.15%)	0 (0.00%)	28,589 (6.42%)	40,494 (9.09%)	13,190 (2.96%)
Process-OOD	1,887	47 (2.49%)	19 (1.01%)	80 (4.24%)	7 (0.37%)	149 (7.90%)
Source-OOD	1,110	17 (1.53%)	3 (0.27%)	59 (5.32%)	3 (0.27%)	59 (5.32%)

Table 13: Sensitivity of Missing labels to local threshold perturbations. "pp" denotes percentage points.

Split	Analyzed	Default support	Support range	Max support change	Max flip rate	Min Jaccard
ID	98,851	10,191 (10.309%)	[9,913, 10,684]	0.499 pp	0.499%	0.9539
Mask-OOD	445,499	40,494 (9.090%)	[40,297, 40,500]	0.044 pp	0.044%	0.9951
Process-OOD	1,887	7 (0.371%)	[2, 10]	0.265 pp	0.265%	0.2857
Source-OOD	1,110	3 (0.270%)	[0, 12]	0.811 pp	0.901%	0.0000

in the macro average, with undefined terms set to zero. The class-support counts below provide the context needed to interpret those main-table scores.

Table 12 gives the corresponding class support. The Mask-OOD count is $445{,}499$ here because this table counts samples with available defect labels; the dataset composition table continues to report all $445{,}500$ Mask-OOD samples. The sparse positive support in Process-OOD and Source-OOD makes their aggregate defect scores sensitive to a small number of cases and should be interpreted cautiously.

G.4 Sensitivity of the threshold-dependent defect rule

Among the defect rules in Appendix E.1, Missing is the only category governed by tunable connected-component thresholds; the other labels use discrete binary topological conditions. We vary the gap threshold over $\{0.015, 0.020, 0.025\}$ and the image/top-component thresholds over $\{0.88, 0.90, 0.92\}$, covering all 27 joint combinations. Table 13 shows that fewer than 0.5% of ID labels and 0.05% of Mask-OOD labels change under any tested setting. The lower Jaccard values on Process-OOD and Source-OOD reflect their very small positive supports rather than large absolute flip rates.

G.5 Additional dataset and external-benchmark checks

The TestPattern topology distribution is 66.7% 1D, 10.8% 1.5D, and 22.5% 2D. Process-OOD contains 17 conditions near the boundary of the usable process window; farther extrapolation produces mostly saturated missing or washout cases. Source-OOD is a controlled transfer test to one held-out Quasar source family, with the ID mask distribution and near-center process conditions kept fixed. It should not be interpreted as broad generalization across multiple unseen source families.

Table 14 compares the common baselines with published LithoSim contour-IoU results. Because the datasets differ, this is supporting evidence rather than a controlled re-ranking experiment. The relatively similar contour-IoU values conceal much larger differences in FabBench Mask-OOD CD error, supporting the use of local metrology alongside overlap metrics.

Table 14: External comparison of contour overlap on LithoSim-Metal and local CD error on FabBench Mask-OOD. CD is reported in nanometers.

Model	LithoSim IoU ↑	FabBench CD ↓
CNN	92.32	24.22
Transformer	91.69	41.55
RFNO	88.58	21.33
MFNO	92.18	19.05

H Benchmark model architectures and training details

H.1 Deterministic Regression Models

Table 15 summarizes the exact code paths and key hyperparameters used for the deterministic baselines.

The CNN and Trans variants keep the same lithography-conditioning block as the other deterministic models, but add a residual skip from the mask embedding to the conditioned latent before decoding. This is the key difference from the original LithoSim-style formulation.

H.2 Physics-Informed Neural Operators

RFNO RFNO represents a conservative approach to spectral modeling, primarily focusing on computational tractability and noise regularization. In this architecture, W is implemented as a local, learnable complex tensor that operates under a strict frequency truncation regime. By limiting the learnable modes to $|k| \leq m$, RFNO effectively applies a low-pass filter to the latent features. While this ensures the capture of coarse-grained structural information and maintains high efficiency, the hard truncation of high-frequency components inherently limits the model's ability to reconstruct sharp intensity gradients and sub-wavelength features critical in advanced lithography.

CFNO CFNO extends the spectral transformation by introducing a patched global linear mechanism. Unlike the point-wise multiplication used in standard FNO, CFNO treats the spectral patches as coupled entities. The weight W is structured as a global complex-valued linear layer acting upon these patches, which enables the model to learn cross-frequency interactions. This configuration facilitates a more holistic transformation of the spectrum, allowing the network to capture complex inter-dependencies between different diffraction orders. Consequently, CFNO offers greater expressivity than RFNO, albeit at the cost of increased parametric complexity.

As shown in Appendix I, CFNO exhibits a qualitatively different failure mode from the other two models. Despite its increased expressivity, CFNO fails to produce physically meaningful SEM images in our setting and frequently diverges during training or inference. This instability arises from its global cross-frequency coupling, which implicitly enforces dense interactions among diffraction orders—an assumption that contradicts the localized and source-constrained nature of the lithographic optical transfer function. As a result, weak and noisy high-frequency components are uncontrollably amplified, leading to severe non-physical oscillations and, in many cases, complete breakdown of image generation. Consequently, CFNO cannot be reliably evaluated on fabrication fidelity metrics such as Boundary IoU and EPE.

MFNO MFNO diverges from the aforementioned models by shifting the partitioning strategy from the frequency domain to the spatial domain. Rather than performing operations on frequency chunks, MFNO utilizes a spatial chunking mechanism. By allowing for localized spatial patches and corresponding learnable adjustments within specific spectral representations, MFNO effectively mimics the localized nature of the optical transfer function in lithographic systems. This spatial-spectral alignment enables the model to preserve critical high-frequency details that are often tied to specific local geometries on the mask.

For consistency across the neural operator frameworks, the encoder and decoder structures within the physics-informed neural operators are designed to be identical to those of our modified CNN, ensuring a fair comparison of their spectral and spatial processing capabilities.

Table 15: **Deterministic image-to-image baselines used in FabBench.**

Model	Backbone	Process conditioning	Role in benchmark
CNN	U-Net-style encoder–decoder with residual feature fusion.	Mask features are fused with source, defocus, and exposure embeddings through cross-attention.	Local convolutional baseline for paired SEM-domain reconstruction.
Trans	ViT encoder with CNN decoder and additional down/up-sampling around the latent bottleneck.	Patch-level mask tokens interact with source and scalar process embeddings through cross-attention.	Transformer-based baseline for modeling longer-range layout–process interactions.

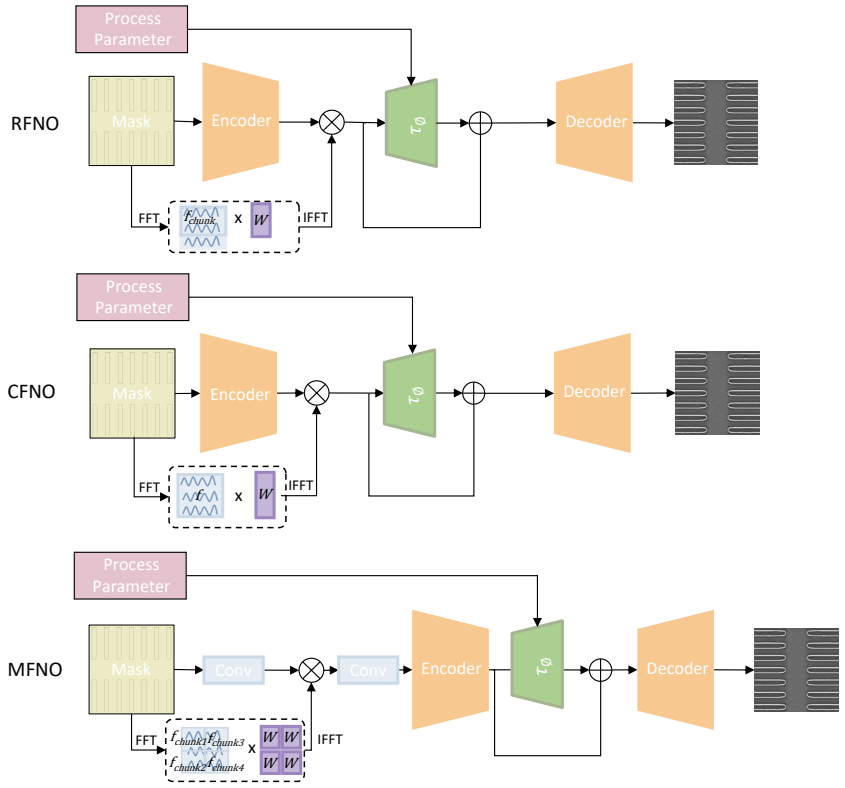


Figure 21: Architecture of physics-informed models.

H.3 Generative Models

We benchmark both GAN-based and diffusion-based generative models to study how different process-conditioning mechanisms affect SEM-domain prediction. All generative variants use the same input tuple (M, S, F, E) and the same SEM-domain supervision, but differ in how source illumination, defocus, and exposure are represented and injected into the generation backbone. This comparison isolates the effect of condition encoding and fusion design under a unified FabBench protocol.

Pix2Pix. Pix2Pix serves as the conditional GAN baseline for paired mask/process-to-SEM translation. Its generator is a skip-connected U-Net, while the discriminator follows the standard patch-level adversarial formulation. To make Pix2Pix process-conditioned, we use early fusion: the binary mask is concatenated with spatially expanded process-condition channels derived from the source, defocus, and exposure inputs. The source condition is represented as a structured 2D source map, while defocus and exposure are normalized using the training ranges and broadcast to constant spatial channels. The generator therefore receives a single multi-channel tensor containing both layout geometry and process information, and predicts the corresponding rendered SEM-domain image. Unlike SD-Adapter, Pix2Pix does not explicitly separate the optical role of the source from the scalar effects of defocus and exposure; all process variables are fused at the input level. This makes

Table 16: **Physics-informed operator baselines used in FabBench.**

Model	Operator design	Lithography conditioning	Benchmark status
RFNO	Low-pass Fourier operator with spectral modes `modes1=modes2=4` and `out_channels=64`.	Mask features are fused with process embeddings through the shared lithography attention block; `rfno_pool=16`.	Included in all benchmark splits.
MFNO	Multi-scale Fourier operator with a shallow mask encoder $1 \rightarrow 8$ and channel multipliers `[16,32]`.	Chunked MFNO blocks fuse spatial mask features with process-conditioned representations.	Included in all benchmark splits.
CFNO	Chunked Fourier operator at 512×512 resolution with `z_channels=64` and `embed_channels=112`.	Concatenates CFNO features before process-aware fusion.	Reported as a diagnostic baseline; current runs are unstable for full quantitative comparison.

Table 17: **Generative models' conditioning variants.** All generative baselines follow the same FabBench input-output protocol.

Model	Encoding	Fusion	Role separation
Pix2Pix	Skip-connected U-Net	Concatenated with mask.	No
SD-AdaIN/SPADE	Entangled process vector	Scale/shift modulation	No
SD-Adapter	Dual cross-attention	Source-conditioned AdaGN	Yes

Pix2Pix a simple and strong paired image-to-image generative baseline, reported as **Pix2Pix** in the main benchmark tables.

SD-AdaIN/SPADE. SD-AdaIN/SPADE encodes the process variables into a single entangled condition vector by concatenating source, exposure, and defocus representations. This vector is then used to generate scale and shift parameters for AdaIN/SPADE-style feature modulation. Although this provides a direct continuous conditioning pathway, it does not explicitly separate the distinct physical roles of the process variables. In particular, the optical source, exposure dose, and defocus are fused before modulation, which may limit the model's ability to distinguish optical image-formation effects from scalar process perturbations.

SD-Adapter. SD-Adapter uses a physics-structured conditioning mechanism based on the distinct roles of lithography process variables. Source illumination primarily determines the optical image-formation basis, while exposure and defocus act as scalar perturbations that modulate thresholding, edge placement, and focus-dependent blur. Accordingly, SD-Adapter extracts process features using dual cross-attention and maps the resulting process tokens to different roles in adaptive group normalization. Specifically, the source condition determines the normalized feature basis through $\mathrm{GN}(f_c)$, while exposure and defocus generate adaptive scale and shift parameters (γ, β) that modulate the normalized features:

$$\mathrm{AdaGN}(f_c; S, F, E) = \gamma(F, E) \cdot \mathrm{GN}_S(f_c) + \beta(F, E), \tag{18}$$

where $\mathrm{GN}_S(\cdot)$ denotes source-conditioned normalization and (γ, β) are derived from exposure and defocus tokens. This design functionally disentangles optical-source effects from scalar process variations, yielding a structured modulation scheme for process-conditioned SEM generation.

Uncertainty of SD-Adapter Unlike deterministic networks that may produce overconfident errors, diffusion architectures enable native uncertainty quantification through repeated stochastic sampling. For each mask and process condition, we run SD-Adapter with 100 independent denoising trajectories using different random seeds and obtain an ensemble of SEM predictions. We compute uncertainty on the predicted SEM intensity field: for each pixel, the 90% prediction interval is formed by the empirical 5th and 95th percentiles across the ensemble. Mean Prediction Interval Width (MPIW@90) is then computed as the spatial average of the interval widths, and Prediction Interval Coverage Probability (PICP@90) is computed as the fraction of target pixels whose benchmark SEM intensity falls inside the corresponding empirical interval. The heatmaps in Figure 23 report these quantities after averaging over all samples at each defocus–exposure condition. These intervals are empirical Monte Carlo intervals and are not post-hoc calibrated. When parameters are pushed to their physical limits, where patterns naturally fail to form, SD-Adapter's MPIW increases

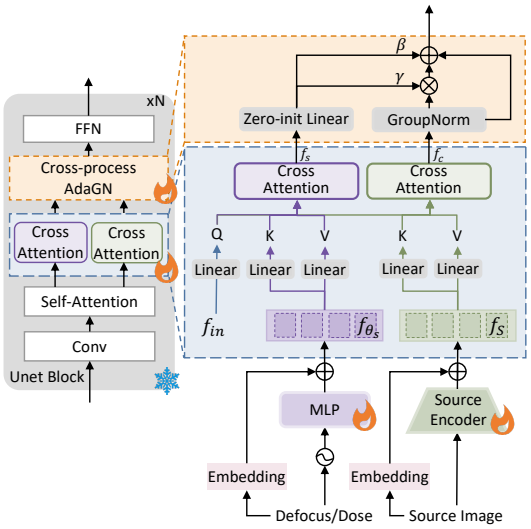
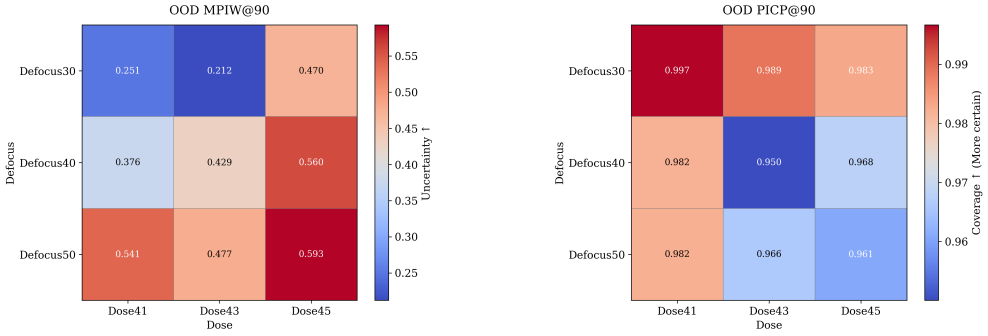


Figure 22: **Overview of SD-Adapter.** SD-Adapter injects lithography process conditions into the diffusion backbone through a structured conditioning pathway. Source illumination defines the source-conditioned normalized feature basis, while exposure and defocus modulate the features through adaptive scale and shift parameters.



(a) OOD MPIW@90 (b) OOD PICP@90

Figure 23: Out-of-distribution (OOD) prediction performance heatmaps under different defocus and exposure dose conditions. (a) Mean Prediction Interval Width (MPIW) at a 90% confidence level, where redder colors and higher values indicate wider prediction intervals and higher predictive uncertainty. (b) Prediction Interval Coverage Probability (PICP) at a 90% confidence level, where redder colors and higher values indicate higher empirical coverage. PICP should be interpreted together with MPIW, since high coverage may also result from wider intervals rather than greater certainty.

substantially, as shown in Figure 23. This suggests that diffusion-sampling variance can provide an uncertainty indicator for severe manufacturing failures that standard deterministic models may fail to capture.

H.4 Implementation Details

All models are trained and evaluated using the same FabBench splits, input-output interface, and evaluator. Each sample contains a mask layout M, source illumination S, defocus F, exposure dose E, and the SEM-domain benchmark target I_{SEM}. Mask and SEM images are resized to a fixed resolution of 512×512 and normalized to $[-1, 1]$. The source is provided as a structured 2D condition, while defocus and exposure are normalized as scalar process variables using the training range. The

same normalization is applied to IID and OOD samples. Model-specific architectural choices and conditioning mechanisms are described in Appendix H. For the deterministic regression models and neural-operator baselines, the optimizer is Adam with weight decay 10^{-5}. The initial learning rate is 10^{-4} for CNN, Trans, and MFNO, and 10^{-5} for RFNO and CFNO. All of these models use the same `ReduceLROnPlateau` scheduler with factor 0.1 and patience 10. The deterministic training configuration uses a maximum of 100 epochs, validation every epoch, and model selection based on validation loss; the callback configuration monitors `val/loss`, saves the last checkpoint, and enables early stopping with patience 5.

For the diffusion baselines, the training learning rate is 10^{-5} and the trainer configuration uses a maximum of 100 epochs with validation every epoch. The diffusion trainer uses full precision (`fp32`), DDP, and gradient accumulation of 1.

All reported experiments in this study were run on 4 NVIDIA A800 GPUs. During evaluation, all prediction results are assessed with the same evaluator. For contour-, defect-, and metrology-related metrics, we apply the same fine-tuned SAM-based contour extraction pipeline to both predictions and targets. CD and EPE are then measured from the extracted top and bottom contours using the predefined gauge annotations.

I ADDITIONAL QUALITATIVE RESULTS

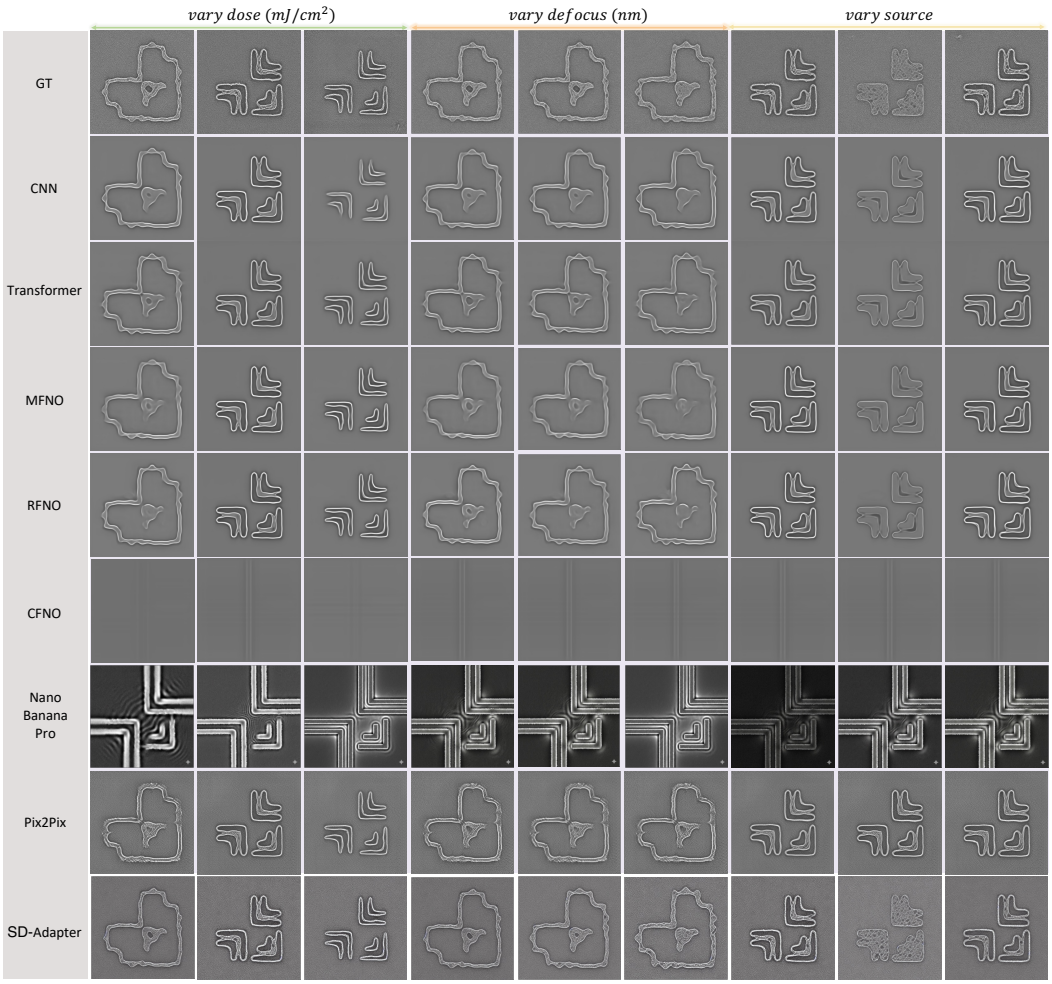


Figure 24: Visual results of case 1 models on *MetalSet*.



Figure 25: Visual results of case 1 models on *MetalSet*.



Figure 26: Visual results of case 2 models on *MetalSet*.

J BROADER IMPACTS

Impact statement. FabBench has potential positive impact by making process-conditioned lithography-twin evaluation more transparent, reproducible, and metrology-aware. In particular, it can help researchers identify models that look visually plausible but fail on CD, EPE, defect, or OOD robustness metrics, which may reduce unnecessary simulation or wafer-trial iterations during early-stage research. The main negative impact is overinterpretation: because FabBench uses SEM-emulated targets and controlled DUV process settings, strong benchmark performance could be misused as evidence of deployment-ready fab accuracy. To mitigate this risk, we explicitly scope FabBench as a research benchmark and exclude raw proprietary SEM data, fab recipes, calibration parameters, and sensitive industrial layouts from release.

FabBench is intended to support open and reproducible evaluation of process-conditioned lithography twins. By providing structured SEM-domain targets, fabrication-aligned metrics, and disentangled OOD splits, the benchmark may help researchers diagnose failure modes that are not captured by generic image similarity metrics. This can support the development of more reliable data-driven lithography models and may reduce unnecessary physical simulation or trial-and-error experimentation during early-stage process exploration.

At the same time, lithography is a high-impact manufacturing technology, and benchmark results should not be interpreted as evidence of deployment-ready fab accuracy. FabBench uses SEM-emulated targets and is scoped to controlled, single-exposure DUV optical lithography settings. Incorrect use of the benchmark as a substitute for wafer-measured SEM validation could lead to overconfidence in model predictions. We therefore emphasize that FabBench is intended as a controlled research benchmark and stress test, not as a production qualification tool.

The released benchmark will exclude raw proprietary SEM images, fab recipes, equipment calibration parameters, and sensitive industrial layouts. We plan to release only SEM-like benchmark targets, anonymized metadata, annotations, splits, evaluation scripts, and baseline configurations that are sufficient for research reproducibility while avoiding disclosure of confidential fabrication information. Since improved lithography modeling can contribute to more efficient semiconductor design and manufacturing, users should ensure that any downstream use complies with applicable institutional, industrial, and export-control policies.

K LIMITATIONS AND FUTURE WORK

FabBench is a SEM-emulated benchmark rather than a substitute for wafer-measured SEM validation. Although its targets are generated through a physics-grounded rendering pipeline with an SEM emulator trained on real SEM images, they should not be interpreted as calibrated ground truth for a specific fabrication line. Strong FabBench performance therefore does not by itself imply deployment-ready fab accuracy.

The current benchmark is also scoped to single-exposure DUV optical lithography with controlled mask layouts, illumination families, and focus–exposure ranges. It does not cover all process knobs, resist stacks, scanner settings, technology nodes, or downstream manufacturing effects such as etch transfer, overlay variation, and multi-patterning. Finally, to provide a reproducible evaluation protocol, FabBench's metrology-aware metric suite relies on contour extraction, predefined gauge locations, and rule-based defect labeling. These choices support consistent comparison across models, but they may not capture all expert-defined criteria used in production metrology.

The following subsections expand these limitations in terms of dataset scale, real-fab coverage, process-variable scope, SEM-emulated targets, and generative-model inference cost, and then outline future directions.

K.1 DATASET DIVERSITY AND SCALE

FabBench combines curated optical stress patterns with synthetic layouts that follow realistic design statistics. This design allows the benchmark to cover optical proximity effects, critical width/spacing combinations, and common local layout structures. However, compared with proprietary industrial databases, the current dataset scale remains limited. Industrial datasets may contain millions of

SEM micrographs and span multiple products, technology nodes, scanner settings, resist stacks, and metrology recipes. FabBench is designed as a public, controlled, and reproducible benchmark, and therefore cannot fully cover all real manufacturing scenarios.

This limitation may affect model generalization in two ways. First, models can learn stable mask-to-SEM mappings for common pattern structures, but may degrade on unseen irregular logic cells, extreme pattern densities, rare layout topologies, or complex local routing patterns. Second, rare process-window corner cases and low-frequency manufacturing defects remain underrepresented, such as local collapse, complex bridging, washout, or failures caused by coupled process effects. Therefore, FabBench should be viewed as a reproducible stress-test benchmark rather than a replacement for industrial-scale fab databases.

K.2 SEM-Emulated Targets and Real-Fab Validation

FabBench constructs benchmark targets using physics-grounded lithography simulation followed by a real-SEM-trained emulator. This design enables dense sampling over source, defocus, and exposure while keeping the benchmark controllable, reproducible, and shareable. However, the targets are SEM-emulated observations rather than wafer-measured SEM ground truth at every process condition.

Consequently, strong performance on FabBench should not be interpreted as deployment-ready accuracy on a specific fabrication line. The SEM emulator introduces real SEM appearance cues, such as white-edge response, contrast variation, and stochastic texture, but it may still retain rendering bias. In particular, when the imaging tool, metrology recipe, resist stack, or inspection condition changes, a domain gap may remain between emulated and measured SEM observations. In addition, FabBench relies on a shared SAM-based contour extractor for downstream CD, EPE, and defect evaluation, so residual contour-extraction errors remain a source of measurement error even though the shared evaluator improves consistency across methods. The goal of FabBench is therefore to provide a controlled and fabrication-aligned evaluation proxy for comparing whether models preserve geometry, metrology behavior, and defect trends under structured variations, rather than to replace wafer-measured SEM validation.

Future versions of FabBench will incorporate more measured wafer SEM samples for external validation, enabling stronger assessment of whether model rankings on SEM-emulated targets correlate with real fab-side behavior. We also plan to study emulator calibration and domain adaptation methods so that the SEM emulator can be efficiently recalibrated for different inspection processes or imaging recipes.

K.3 Scope of Lithography Settings and Process Variables

FabBench currently focuses on single-exposure DUV optical lithography with controlled variations in source illumination, defocus, and exposure. This setting supports controlled evaluation of process-conditioned SEM prediction, but it does not cover all lithography scenarios.

First, FabBench does not model multiple-exposure or multi-patterning processes. In advanced manufacturing, the final printed structure may depend on interactions among multiple exposures, overlay errors, etch transfer, and downstream process integration. These multi-step patterning effects are outside the current benchmark scope.

Second, FabBench does not claim cross-tool or cross-equipment generalization. The SEM emulator is trained under a specific imaging regime, and the lithography simulator follows a fixed modeling configuration. Although we include a small cross-domain adaptation study for a neighboring inspection setting, performance on FabBench should not be interpreted as evidence of direct transfer across scanners, SEM tools, metrology recipes, resist stacks, or fabrication lines without recalibration.

Third, FabBench should not be interpreted as an EUV lithography benchmark. EUV lithography involves distinct physical mechanisms, including reflective mask optics, different illumination and imaging characteristics, stochastic photon effects, and EUV-specific resist behavior. Conclusions drawn from FabBench therefore do not automatically transfer to EUV process modeling.

The current benchmark focuses on source illumination, defocus, and exposure because these variables are central to DUV process-window behavior. Other manufacturing factors, including resist

chemistry, etch effects, mask error, overlay variation, scanner calibration, wafer topography, and stochastic wafer-level variations, remain outside the current scope and are left for future extensions.

K.4 Future Work

Based on the above limitations, we identify several directions for future development.

Larger and more diverse layout coverage. Future versions will include larger and more realistic layout distributions, such as complex logic cells, standard-cell combinations, dense routing patterns, contact/via structures, and multi-layer design contexts. Broader layout coverage will help evaluate whether models learn transferable lithography responses rather than memorizing a limited set of local pattern primitives.

Measured SEM validation and emulator calibration. Future work will incorporate more measured wafer SEM samples to validate whether SEM-emulated benchmark performance correlates with real fab-side metrology behavior. We will also investigate emulator recalibration, domain adaptation, and inspection-recipe adaptation to support benchmark extension across different tools, imaging conditions, and fabrication settings.

Expansion of process variables. Beyond source, defocus, and exposure, future versions of FabBench can include resist-related parameters, etch bias, mask error, overlay variation, scanner settings, and stochastic wafer-level effects. This would extend FabBench from optical process-window evaluation toward broader manufacturing-aware lithography twin evaluation.

Improving precision and physical consistency. Future models need to satisfy both SEM realism and metrology precision. We plan to explore hybrid architectures that combine physics-informed constraints, neural operators, or deterministic geometric predictors with diffusion-based generative models. One possible direction is to use a deterministic branch to constrain contours, CD, and EPE, while a generative branch synthesizes SEM texture and stochastic appearance. This could reduce geometric error while preserving realistic observation-domain details.

Acceleration of generative lithography twins. To improve industrial usability, future work will further reduce the inference cost of generative models. In addition to distillation, possible directions include model pruning, adapter compression, shared latent caching, process-conditioned feature reuse, and multi-resolution generation. These techniques could support larger process-window sweeps and near-real-time inspection assistance.

Overall, FabBench is designed as a public, controlled, and metrology-aware benchmark. By incorporating more measured SEM validation, broader process variables, larger layout distributions, and more efficient model architectures, future versions can further support the development and evaluation of fabrication-aware lithography twins.