HOIST: A High-Order Interaction Surrogate for Sample-Efficient High-Sigma Yield Estimation in High-Dimensional Circuits

Zhongru Xiong[1,2], Mingjun Wang[3,*], Yanfang Liu[3], Tsung-Yi Ho[3], Bei Yu[2,3], Binwu Zhu[1,*]
[1]Southeast University [2]Primarius Technologies [3]The Chinese University of Hong Kong

Abstract

Accurate high-sigma yield estimation requires resolving rare circuit failures in high-dimensional process-variation spaces, where exhaustive transistor-level Monte Carlo (MC) simulation is prohibitively expensive. Surrogates can screen large populations cheaply, but flat models do not explicitly exploit device organization, and directly counting predicted failures can turn small threshold errors into large probability bias. We present HOIST, a device-structured and SPICE-validated framework that separates failure discovery from probability estimation. A shared device encoder, global circuit context, and multiplicative local–global fusion provide an explicit pathway for coordinated mismatch. An always-on attention residual head further refines high-risk ordering with a fixed label budget and a strictly positive calibration scale. The frozen models process a finite MC population in two bounded-memory passes, retaining complementary failure-prone and boundary-sensitive candidates for direct validation. Fixed score strata and fresh uniform audits provide an unbiased fallback for failures missed outside Top-K; model predictions are not final failure labels. Experiments on SRAM benchmarks with 18–1203 process variables evaluate failure-probability error and total transistor-level simulation cost against representative high-sigma yield-analysis baselines. HOIST requires 2,848–5,336 SPICE simulations with 2.59%–8.28% relative failure-probability error and achieves average speedups over six baselines of 3.76×, 7.73×, and 6.64× on three cases, respectively (6.04× overall).

1 Introduction

Process variation is a first-order signoff constraint in scaled integrated circuits because mismatch perturbs nominally identical devices [1]. In SRAM, intrinsic fluctuations degrade cell stability and make array-level failure probability a central design concern [2, 3]. Modern signoff may target per-instance failure probabilities near 10^{-5} while involving hundreds or thousands of process variables, making accurate transistor-level estimation prohibitively expensive [4–6]. Throughout this paper, *high-sigma yield estimation* denotes the circuit task of estimating a very small parametric failure probability, whereas *rare-event probability estimation* refers to its broader statistical setting.

Direct MC is unbiased but requires roughly 10^7 transistor-level simulations to estimate a failure probability of one in 100,000 with 10% relative standard error [7–9]. Importance sampling draws and reweights failure-prone samples, while surrogate and hybrid methods use learned responses to guide costly physical evaluations [4, 5, 10–15]. Despite this progress, high-dimensional circuits still pose two coupled challenges: discovering coordinated failure patterns and converting model-guided discovery into a valid probability estimate.

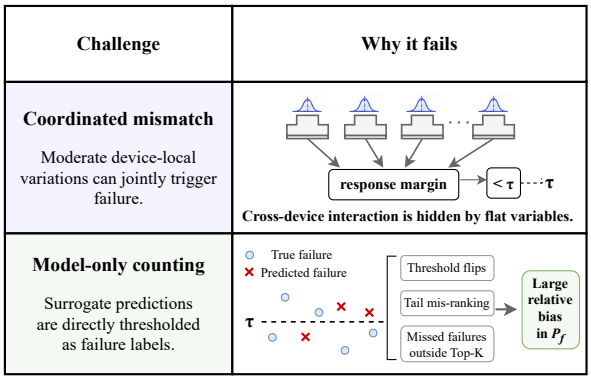

Figure 1: Motivation for HOIST: device-structured discovery captures coordinated mismatch, while SPICE-validated stratified estimation avoids model-only counting bias.

Figure 1 summarizes these challenges and the corresponding HOIST responses. The first is *failure discovery under coordinated mismatch*: several moderate device-local deviations can jointly reduce read margin, increase delay, or create an offset even when no variable is extreme [2]. Mean-shift proposals may under-cover curved or disconnected failure regions, while global low-rank approximations can suppress weak variables with strong joint effects [9, 16]. With limited labels, a flat representation does not encode shared processing within each transistor or conditioning across devices; structured learning provides this useful inductive bias [17].

The second challenge is *turning discovery into a valid probability estimate*. Low average regression error does not prevent mis-ranking the few samples that determine P_f, and small response errors near the specification can flip enough labels to dominate a 10^{-5} estimate. Validating only a high-risk prefix removes false positives within it but leaves false negatives outside; treating all unvalidated samples as passing therefore gives a lower bound. Predictions should instead concentrate the SPICE budget, while a sampling design with known inclusion probabilities determines final probability and uncertainty.

These observations motivate two complementary priors. A MOS device provides a natural local unit for shared encoding, while circuit context modulates each local representation to expose cross-device interactions. The surrogate score then partitions a finite MC population into risk strata: high-risk candidates receive direct validation, and randomized audits in the remaining strata prevent silent exclusion. A better surrogate can thus reduce cost without turning its errors into final failure labels.

We propose HOIST, a device-structured, threshold-aware, and SPICE-validated framework for high-sigma yield estimation. From a limited DoE, its shared encoder, circuit context, multiplicative fusion, and threshold-aware weighting learn a continuous response. The frozen surrogate then screens a finite MC stream; complementary heaps and fixed-budget residual calibration refine candidates for

*Corresponding authors.

direct validation, with a positive scale selected from a predeclared set. A second pass freezes score strata and uses pilot-guided fresh uniform audits, providing a design-unbiased fallback for failures outside the retained Top-K set.

The contributions of this paper are summarized as follows:

- We introduce a device-structured interaction surrogate that combines shared local encoding, global circuit context, multiplicative fusion, and threshold-aware training to model coordinated mismatch.
- We develop a bounded-memory two-pass discovery flow with complementary Top-K retention and budget-controlled residual calibration using disjoint fitting, selection, and diagnostic labels.
- We combine frozen score strata with randomized SPICE audits to obtain a design-unbiased finite-population estimate without treating surrogate predictions as failure labels.
- Across SRAM benchmarks with 18–1203 variables, HOIST achieves 2.59%–8.28% relative error using 2,848–5,336 SPICE simulations, with the lowest end-to-end CPU runtime, 26%–68% below the fastest competing method. Across five Case-18 seeds, it completes all trials without any relative error exceeding 20%.

2 Background and Related Work

High-Sigma Yield Estimation. For a circuit subject to process variation, parametric yield is conventionally defined as the probability that its performance satisfies the prescribed specification; the complementary probability of a specification violation is the failure probability P_f [5, 10]. High-sigma yield estimation concerns the accurate estimation of a very small P_f, as required when a large circuit contains many replicated or variation-sensitive components. Direct Monte Carlo (MC) is unbiased, but observing enough rare failures requires a simulation count that grows inversely with P_f. Because every sample entails a transistor-level circuit evaluation, accurate high-sigma estimation can therefore incur a prohibitive physical-simulation cost [4, 5, 10].

Surrogate and Hybrid Yield Analysis. Surrogate models reduce simulation cost by approximating a circuit response or failure boundary from a limited set of physical labels. Low-rank tensor models, high-dimensional Bayesian optimization, and shrinkage deep-kernel methods show that inexpensive predictions can guide the search for informative circuit variations [5, 14, 18]. Their effectiveness, however, depends on tail fidelity rather than average response accuracy alone. A small regression error near a specification can reverse a binary failure label, and a small number of such reversals can dominate an estimate of a small failure probability.

This distinction is especially important in circuits with many device-local variations. A response can depend on coordinated deviations across several devices even when no individual variable is extreme. Generic global regressors may represent these interactions only after observing sufficient tail labels, whereas low-complexity approximations can suppress weak individual effects whose joint contribution is large. These difficulties motivate evaluation criteria that measure both tail-estimation accuracy and the complete physical-simulation cost rather than a global fitting error alone.

Hybrid methods incorporate learned models into the IS pipeline. OPTIMIS combines failure-aware sampling with a normalizing flow to learn an expressive proposal distribution, while FUSIS explicitly fuses surrogate modeling with IS [4, 15]. In these methods, the learned model assists failure discovery or proposal construction, whereas probability estimation typically retains physical evaluations and likelihood-ratio weighting. This differs from directly thresholding surrogate predictions, but still requires adequate proposal support.

Model-Assisted Sampling and Validation. Model-assisted sampling uses a prediction score to allocate expensive labels, while a randomized sampling design supplies the observations used for inference. In finite-population stratified sampling, a population is partitioned into fixed strata, samples are selected with known inclusion probabilities, and stratum means are weighted by stratum sizes [19]. This separation is useful because a score can be highly informative for prioritization without being a calibrated failure probability. Unlike surrogate–IS hybrids, the score need not parameterize a proposal density or enter a likelihood ratio; it can be frozen and used solely for validation-budget allocation, leaving probability estimation to randomized physical observations with known inclusion probabilities.

The distinction also exposes a common statistical risk. If a model deterministically selects a high-risk prefix and every unselected sample is treated as passing, undetected failures create downward bias. Randomized audits protect against silent exclusion only when the strata and final-audit sample sizes are fixed before final-audit labels are observed, and when every non-census region retains a known nonzero chance of inspection. Consequently, low-cost discovery and valid probability estimation must be assessed as related but distinct objectives.

3 Problem Formulation

Let $\boldsymbol{x} \in \mathbb{R}^d$ denote a normalized process-variation vector, where d is the number of process variables. Its components are modeled as mutually independent standard Gaussian variables [4, 5],

$$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d), \tag{1}$$

For a given $\boldsymbol{x}$, the transistor-level simulator is treated as an expensive black-box function that returns a scalar circuit performance metric,

$$y = f_{\mathrm{SPICE}}(\boldsymbol{x}). \tag{2}$$

All benchmarks in this work use a larger-is-worse specification with threshold τ. The physical failure indicator is defined as

$$I(\boldsymbol{x}) = \begin{cases} 1, & f_{\mathrm{SPICE}}(\boldsymbol{x}) > \tau, \\ 0, & f_{\mathrm{SPICE}}(\boldsymbol{x}) \le \tau. \end{cases} \tag{3}$$

The circuit failure probability and parametric yield are

$$P_f = \Pr[f_{\mathrm{SPICE}}(\boldsymbol{x}) > \tau] = \int_{\mathbb{R}^d} I(\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x}, \qquad Y = 1 - P_f, \tag{4}$$

where $p(\boldsymbol{x})$ is the probability density of the process variations. Direct Monte Carlo draws N independent samples $\{\boldsymbol{x}_i\}_{i=1}^N$ from $p(\boldsymbol{x})$ and estimates the failure probability by the observed failure fraction,

$$\widehat{P}_f^{\mathrm{MC}} = \frac{1}{N} \sum_{i=1}^{N} I(\boldsymbol{x}_i), \qquad \mathrm{Var}\left(\widehat{P}_f^{\mathrm{MC}}\right) = \frac{P_f(1 - P_f)}{N}. \tag{5}$$

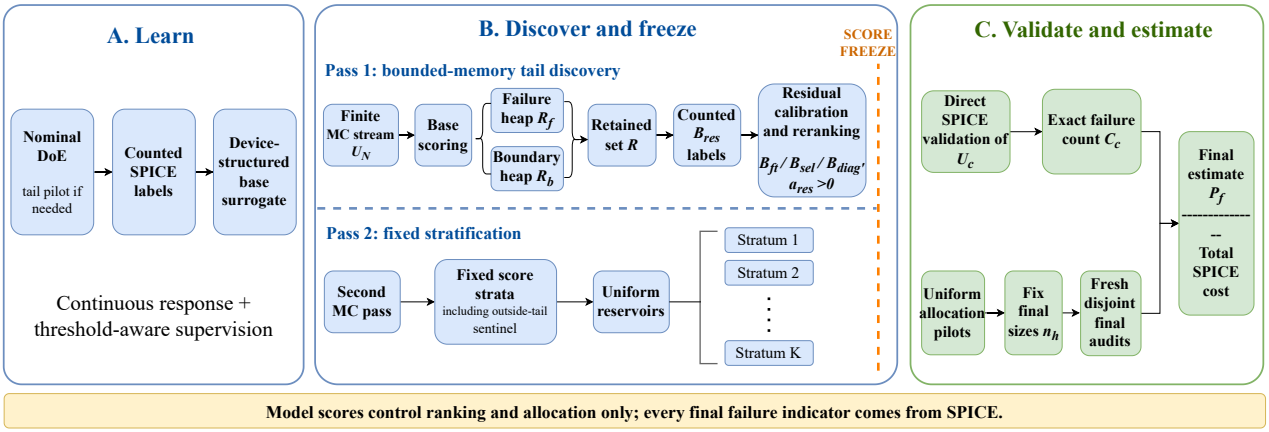

Figure 2: Overview of HOIST. Part A trains the structured surrogate, Part B discovers and reranks tail candidates before freezing score strata, and Part C combines direct validation with randomized audits for final estimation.

Because P_f is very small in high-sigma analysis, direct MC requires a large N and hence a prohibitive number of transistor-level simulations. The objective is to estimate P_f accurately while minimizing the total number of SPICE evaluations.

4 The HOIST Framework

Figure 2 gives the end-to-end HOIST workflow and separates its three roles. Part A learns a continuous device-structured response surrogate from counted physical labels. Part B performs bounded-memory tail discovery, freezes the calibrated score, and makes a second pass to construct fixed audit strata. Part C uses direct validation and fresh randomized SPICE observations to form the finite-population estimate.

4.1 Device-Structured Interaction Surrogate

Device tokenization and shared local encoder. For a circuit with m MOS devices, write the normalized variation vector as $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_m^\top, \boldsymbol{x}_g^\top]^\top$. Here $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]^\top \in \mathbb{R}^{m \times d_\ell}$ collects device-local variables, missing local entries are masked, and $\boldsymbol{x}_g \in \mathbb{R}^{d_g}$ contains any shared chip- or batch-level variables ($d_g = 0$ if absent). A learned NMOS/PMOS embedding $\boldsymbol{e}_{t_j} \in \mathbb{R}^{d_h}$ carries device-type information. Figure 3 illustrates the complete local–global computation. Each MOS device is treated as a natural local unit, while a shared two-layer encoder maps its local description to a device token and pools limited supervision across devices,

$$\boldsymbol{h}_j = \phi_\theta([\boldsymbol{x}_j, \boldsymbol{e}_{t_j}]) \in \mathbb{R}^{d_h}. \tag{6}$$

After fixed padding, $d = md_\ell + d_g$; deterministic masks are metadata rather than additional random variables. The shared vector $\boldsymbol{x}_g$ is represented once rather than copied as independent device noise, and deterministic netlist order preserves device identity.

Global context and multiplicative fusion. Local variation alone cannot identify whether a device is harmful under the current mismatch pattern. We compress every token and construct a global circuit context in deterministic netlist order:

$$\boldsymbol{g} = \Psi_\theta([\kappa(\boldsymbol{h}_1), \ldots, \kappa(\boldsymbol{h}_m), \boldsymbol{x}_g]) \in \mathbb{R}^{d_h}. \tag{7}$$

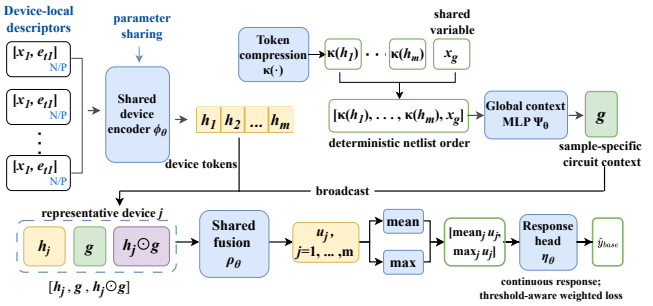

Figure 3: Device-structured local–global surrogate. Shared encoding produces device tokens; ordered compressed tokens and shared variables form context $\boldsymbol{g}$, which is broadcast through $[\boldsymbol{h}_j, \boldsymbol{g}, \boldsymbol{h}_j \odot \boldsymbol{g}]$ before mean/max response pooling.

The upper context branch in Fig. 3 preserves deterministic netlist order and thereby retains the distinct electrical roles of cell, peripheral, and control devices. Equation (7) exposes the complete pattern without explicitly enumerating $O(m^2)$ device pairs. The context is broadcast back to every device, and the response readout uses

$$\begin{aligned} \boldsymbol{u}_j &= \rho_\theta([\boldsymbol{h}_j, \boldsymbol{g}, \boldsymbol{h}_j \odot \boldsymbol{g}]) \in \mathbb{R}^{d_h}, \\ \widehat{y}_{\text{base}}(\boldsymbol{x}) &= \eta_\theta([\text{mean}_j \boldsymbol{u}_j, \max_j \boldsymbol{u}_j]). \end{aligned} \tag{8}$$

As highlighted by the broadcast path in Fig. 3, the product $\boldsymbol{h}_j \odot \boldsymbol{g}$ explicitly conditions each local representation on the circuit mismatch pattern, while the direct paths retain non-interacting information. Mean/max pooling captures distributed and strong localized effects. The shared local encoder pools limited supervision without erasing device identity because $\boldsymbol{g}$ remains sample specific; together, the two mechanisms avoid explicit quadratic enumeration of device pairs.

Threshold-aware response supervision. The surrogate predicts the continuous SPICE response before it is used for failure ranking. For n_{tr} training pairs $(\boldsymbol{x}_i, y_i)$ with $y_i = f_{\text{SPICE}}(\boldsymbol{x}_i)$, let $s \in \{+1, -1\}$ specify the failure direction: $s = +1$ for larger-is-worse and $s = -1$ for smaller-is-worse. All experiments use $s = +1$. Define the signed failure margin as $d_i = s(y_i - \tau)$. Let $\alpha_f, \alpha_b \geq 0$ be loss weights, and

$\sigma_b > 0$ be the boundary bandwidth. The training loss is

$$\mathcal{L}_{\text{base}} = \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \left[1 + \alpha_f \mathbb{I}(d_i > 0) + \alpha_b e^{-|d_i|/\sigma_b}\right] (\widehat{y}_{\text{base}}(\boldsymbol{x}_i) - y_i)^2. \tag{9}$$

Equation (9) increases supervision near the specification and on observed failures while retaining the signed physical margin needed for tail and boundary ranking in Eq. (11). A small counted tail pilot is used only when the initial design lacks tail observations, and every pilot label is included in the reported SPICE ledger.

4.2 Tail Discovery and Budget-Controlled Residual Calibration

Figure 4 details the ordering and label isolation used in Part B. The upper lane performs bounded-memory tail retention, whereas the lower lane uses a predeclared, counted, and disjoint residual-label budget to refine the retained ordering.

Complementary bounded-memory tail retention. Let $\mathcal{U}_N = \{\boldsymbol{x}_i\}_{i=1}^N$, with $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$, denote a finite MC candidate population generated once and held fixed throughout both streaming passes. For a streamed candidate $\boldsymbol{x} \in \mathcal{U}_N$, let $\widehat{d}_{\text{base}}(\boldsymbol{x}) = s(\widehat{y}_{\text{base}}(\boldsymbol{x}) - \tau)$ be its signed predicted margin. The scalar screening score

$$S_{\text{base}}(\boldsymbol{x}) = \lambda_f \widehat{d}_{\text{base}}(\boldsymbol{x}) - \lambda_b |\widehat{d}_{\text{base}}(\boldsymbol{x})| \tag{10}$$

uses fixed nonnegative weights λ_f and λ_b and preserves the failure–boundary trade-off. It supplies downstream diagnostics but does not determine heap membership; HOIST instead maintains complementary fixed-capacity heaps. Let $\text{TopK}_K(\mathcal{A}; f)$ denote the K elements of candidate set $\mathcal{A}$ with largest scalar score f:

$$\begin{aligned} \mathcal{R}_f &= \text{TopK}_{K_f}(\mathcal{U}_N; \widehat{d}_{\text{base}}), \\ \mathcal{R}_b &= \text{TopK}_{K_b}(\mathcal{U}_N; -|\widehat{d}_{\text{base}}|), \\ \mathcal{R} &= \mathcal{R}_f \cup \mathcal{R}_b, \qquad K = |\mathcal{R}|. \end{aligned} \tag{11}$$

As shown in the upper lane of Fig. 4, the failure and boundary heaps are maintained independently and merged only after streaming, preventing a single scalar trade-off from determining retention. The former seeks large positive margins, while the latter preserves threshold-sensitive candidates; duplicates are retained once.

Always-on residual calibration under a counted budget. The base model can exhibit systematic response errors after tail selection. We therefore execute residual calibration in every full HOIST run, with its only adjustable cost fixed before execution as $B_{\text{res}} = B_{\text{fit}} + B_{\text{sel}} + B_{\text{diag}}$. The three branches in the lower lane of Fig. 4 emphasize that residual fitting, scale selection, and ranking-risk diagnosis use disjoint physical labels drawn from $\mathcal{R}$; the base surrogate is frozen. Given fused tokens $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m]^\top \in \mathbb{R}^{m \times d_h}$, a four-head cross-attention module uses the scalar base prediction after a learned query projection Q:

$$\begin{aligned} \boldsymbol{a}(\boldsymbol{x}) &= \text{MHA}(Q(\widehat{y}_{\text{base}}(\boldsymbol{x})), \boldsymbol{U}, \boldsymbol{U}), \\ q_\omega(\boldsymbol{x}) &= \eta_\omega([\boldsymbol{a}(\boldsymbol{x}), \text{mean}_j\, \boldsymbol{u}_j, \max_j \boldsymbol{u}_j, \widehat{y}_{\text{base}}(\boldsymbol{x})]) \in \mathbb{R}. \end{aligned} \tag{12}$$

Thus q_ω predicts a residual in the same normalized response units as $\widehat{y}_{\text{base}}$, rather than a new response from scratch. Its held-out tail-ranking loss selects a strictly positive, predeclared scale:

$$\widehat{y}_{\text{cal}}(\boldsymbol{x}) = \widehat{y}_{\text{base}}(\boldsymbol{x}) + \alpha_{\text{res}} q_\omega(\boldsymbol{x}), \qquad \alpha_{\text{res}} \in \mathcal{A}_\alpha \subset (0, 1]. \tag{13}$$

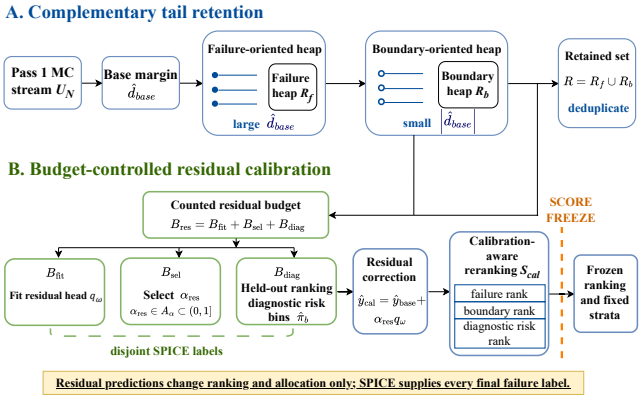

Figure 4: Complementary tail discovery and budget-controlled residual calibration using independent heaps, disjoint labels, and frozen scores before final auditing.

Residual correction is always enabled in full HOIST. Thus, α_{res} calibrates its magnitude over the predeclared strictly positive set $\mathcal{A}_\alpha$; removing the complete stage is considered only in ablation.

Calibration-aware reranking and score freeze. Figure 4 summarizes the separation of model fitting, scale selection, and ranking-risk assessment, which prevents the residual head from relabeling candidates. In held-out calibrated-margin bin b, with v_b labels and $z_b \in \{0, \ldots, v_b\}$ physical failures, define

$$\begin{aligned} \widehat{\pi}_b &= \frac{1 + z_b}{2 + v_b}, \\ S_{\text{cal}}(\boldsymbol{x}) &= \lambda_f R(\widehat{d}_{\text{cal}}(\boldsymbol{x})) + \lambda_b R(-|\widehat{d}_{\text{cal}}(\boldsymbol{x})|) + \lambda_u R(\widehat{\pi}_{b(\boldsymbol{x})}), \end{aligned} \tag{14}$$

where $\widehat{d}_{\text{cal}}(\boldsymbol{x}) = s(\widehat{y}_{\text{cal}}(\boldsymbol{x}) - \tau)$ and R maps a quantity to $[0, 1]$ by ascending rank. Equation (14) reranks $\mathcal{R}$ and defines the score ordering used by the subsequent audit allocation; it never substitutes for a physical failure indicator. After this score is frozen, the remaining finite population is partitioned into fixed score strata, including one global outside-tail sentinel stratum, before any final audit is drawn.

Rank normalization in Eq. (14) prevents a raw margin, boundary distance, or empirical bin risk from dominating because of scale, so the fixed nonnegative weights express only the predeclared discovery trade-off. The smoothed $\widehat{\pi}_b$ stabilizes sparsely labeled bins but is not a calibrated population failure probability.

Two-pass separation and charged decisions. The complementary heaps are constructed from the frozen base model before residual labels are acquired, confining B_{res} to the retained set. The residual head changes only candidate priority and fixed audit allocation; every physical label used for training, calibration, validation, or auditing is charged once. Beyond the freeze boundary in Fig. 4, neither the model nor residual labels may adapt the strata using final-audit outcomes.

4.3 SPICE-Validated Estimation

As shown in Part C of Fig. 2, directly validated census and randomized stratum audits contribute separately to the final estimator.

After score freezing, the retained union $\mathcal{U}_c = \mathcal{R}_f \cup \mathcal{R}_b$ is completely validated by SPICE; let C_c be its failure count. The H fixed strata over the remainder have sizes N_h, where $N = |\mathcal{U}_c| + \sum_{h=1}^H N_h$.

Algorithm 1 HOIST: Structured Tail Discovery and Estimation

Input: B_{res}, N, (K_f, K_b), H, and $\mathcal{A}_\alpha \subset (0, 1]$
Output: $\widehat{P}_{f,N}$ and charged SPICE cost

1: Label the initial design; run a counted training-tail pilot only if needed;
2: Train the device-structured base surrogate using Eq. (9);
3: Stream candidates and retain $\mathcal{R}_f, \mathcal{R}_b$ by Eq. (11);
4: Acquire disjoint $B_{\text{fit}}, B_{\text{sel}}, B_{\text{diag}}$ labels; fit q_ω;
5: Select $\alpha_{\text{res}} \in \mathcal{A}_\alpha$, rerank by Eq. (14), and freeze scores and strata;
6: Directly validate $\mathcal{U}_c = \mathcal{R}_f \cup \mathcal{R}_b$;
7: Draw allocation pilots, fix n_h, then draw fresh disjoint final audits;
8: Compute Eq. (15);
9: **return** the estimate and complete cost ledger;

Table 1: Benchmark Characteristics and Direct-MC References.

Case	Dim.	Threshold	Ref. P_f	MC calls
Case-18	18	$1.731e{-1}$	$5.800e{-5}$	1,000,000
Case-243	243	$1.085e{-5}$	$6.180e{-5}$	1,280,000
Case-1203	1203	$1.171e{-4}$	$5.238e{-5}$	1,680,000

Direct MC values define the error and speedup references used throughout this section.

Each stratum receives a uniform allocation pilot of size r_h, observing a_h failures and fixing n_h within predeclared bounds (8–24). A fresh, disjoint uniform audit then observes z_h failures among n_h draws from the remaining $M_h = N_h - r_h$ candidates. The estimator is

$$\widehat{P}_{f,N} = \frac{1}{N}\left[C_c + \sum_{h=1}^{H}\left(a_h + M_h \frac{z_h}{n_h}\right)\right]. \tag{15}$$

Conditioned on frozen strata and the allocation-pilot outcome, Eq. (15) is design-unbiased over the fresh audit randomization. Surrogate predictions determine only retention, ordering, and stratification; every failure indicator entering Eq. (15) is obtained from SPICE.

5 Experimental Results

5.1 Experimental Setup

Benchmarks and metrics. We evaluate HOIST on three FreePDK45 SRAM yield-estimation benchmarks derived from OpenYield-style netlists [6]. Case-18 models an isolated 6T cell, while Case-243 and Case-1203 include progressively larger peripheral circuits. All process variables are normalized and independently sampled from standard Gaussian distributions. Each benchmark uses a larger-is-worse specification. Table 1 lists the failure-probability estimates obtained by direct MC, which serve as the references. We report the estimated failure probability P_f, its relative error with respect to the corresponding reference, the total number of transistor-level SPICE calls, and the resulting simulation-count speedup over direct MC.

The three cases retain the same evaluation semantics while increasing the number of process variables and circuit context. Thus, the cross-case trend in Tables 3 and 5 isolates scaling behavior without changing the physical failure definition or the cost unit.

Comparison methods and evaluation protocol. We compare against MNIS [10], ACS [16], AIS [11], HSCS [20], LRTA [14], and OPTIMIS [4]. Every reported SPICE total charges all physical simulations used for initial failure discovery, proposal or model construction, validation, and final estimation. For Case-18 baselines evaluated over repeated trials, we report the median successful run; the remaining baseline entries are controlled runs under their method-specific termination criteria, including an FOM target of 0.1 where applicable. This FOM threshold is used only by the corresponding baselines and is not applied to HOIST, whose fixed simulation and audit budgets are specified below. Thus, the comparison uses complete transistor-level simulation cost for every method.

Table 2: HOIST Simulation Budgets and Audit Settings.

Parameter	Case-18	Case-243	Case-1203
Initial labels	1,024	1,024	2,048
B_{res} labels	500	800	1,500
MC population	10^7	10^7	10^7
Failure / boundary heap	1,000 / 200	1,000 / 200	1,000 / 200
Fixed audit strata (incl. sentinel)	8	8	8
Reservoir per score stratum	128	128	128
Pilot samples per stratum	4	4	4
Final samples per stratum	8–24	8–24	8–24

All residual, pilot, validation, and final-audit labels are charged; pilot and final audits are disjoint.

HOIST configuration. HOIST uses 1024 initial labels for Case-18 and Case-243 and 2048 for Case-1203. The surrogate ranks a large MC stream using failure-oriented and boundary-oriented retention, followed by SPICE validation and stratified auditing. The reported SPICE budget includes these initial labels and every residual-calibration label used to train the frozen residual checkpoint. In every full run, residual calibration is executed with $\alpha_{\text{res}} \in \mathcal{A}_\alpha \subset (0, 1]$; its only variable cost is the explicit B_{res} budget. Removing the complete stage is examined solely as an ablation in Table 5.

Table 2 fixes the complete label ledger for the full model. All listed physical labels enter the SPICE totals in Table 3.

The protocol fixes residual splits and score strata before the final-audit labels are obtained. Consequently, the comparison separates costly physical labels from inexpensive surrogate inference and reports the former in a common unit across all methods.

5.2 Yield Estimation Results and Analysis

Case-18. For the low-dimensional cell, HOIST achieves 8.28% relative error with 3,748 SPICE calls, 33% fewer than OPTIMIS, using fixed-budget ranking and direct validation without fitting a sampling proposal.

Case-243. The shared encoder, global context, and multiplicative fusion capture coordinated device behavior with limited labels. HOIST achieves 2.59% relative error with 60% fewer calls than OPTIMIS and less than one third of the LRTA cost.

Case-1203. At the highest dimension, HOIST combines compressed global context, complementary tail retention, residual calibration, and randomized SPICE audits, achieving 4.73% error with 37% fewer calls than OPTIMIS.

5.3 Computational Time Study

Table 4 reports end-to-end CPU time, including SPICE simulation and method overhead. Although MNIS has inexpensive norm-based search at low dimension, HOIST still reduces Case-18 runtime from 0.04 to 0.03 hours. At higher dimensions, OPTIMIS must learn a failure-focused flow, whereas HOIST reuses a parameter-shared device encoder, avoids explicit pairwise interactions, fits the residual only on a fixed retained set, and streams 10^7 candidates through

Table 3: Yield-Estimation Accuracy, SPICE Cost, and Speedup.

Method	Case-18				Case-243				Case-1203			
	P_f	Rel. Err.	#SPICE	Speedup	P_f	Rel. Err.	#SPICE	Speedup	P_f	Rel. Err.	#SPICE	Speedup
MC ref.	$5.80e{-5}$	–	1,000,000	1.00×	$6.18e{-5}$	–	1,280,000	1.00×	$5.24e{-5}$	–	1,680,000	1.00×
MNIS [10]	$5.15e{-5}$	11.24%	32,643	30.63×	$6.03e{-5}$	2.39%	21,687	59.02×	$5.32e{-5}$	1.64%	48,790	34.43×
ACS [16]	$4.92e{-5}$	15.12%	8,457	118.25×	$5.40e{-5}$	12.59%	44,230	28.94×	$6.25e{-5}$	19.22%	59,392	28.29×
AIS [11]	$5.38e{-5}$	7.33%	12,496	80.03×	$5.16e{-5}$	16.47%	36,550	35.02×	$4.35e{-5}$	16.88%	37,658	44.61×
HSCS [20]	$5.42e{-5}$	**6.48%**	17,118	58.42×	$5.58e{-5}$	9.77%	13,538	94.55×	$5.88e{-5}$	12.33%	35,840	46.88×
LRTA [14]	$5.07e{-5}$	12.59%	8,338	119.93×	$6.07e{-5}$	**1.80%**	8,929	143.35×	$5.45e{-5}$	4.05%	22,300	75.34×
OPTIMIS [4]	$5.37e{-5}$	7.38%	5,600	178.57×	$5.49e{-5}$	11.23%	7,200	177.78×	$5.31e{-5}$	**1.41%**	8,500	197.65×
HOIST	$5.32e{-5}$	8.28%	**3,748**	**266.81×**	$6.02e{-5}$	2.59%	**2,848**	**449.44×**	$4.99e{-5}$	4.73%	**5,336**	**314.84×**

Lower relative error and charged SPICE cost are preferred; speedup is relative to the direct-MC budgets in Table 1.

Table 4: CPU Wall-Clock Runtime Comparison (Hours).

Method	Case-18	Case-243	Case-1203
MC	10.38	113.71	783.47
MNIS	0.04	3.44	22.75
ACS	0.10	3.89	27.70
AIS	0.42	3.28	17.56
HSCS	0.38	1.11	16.71
LRTA	0.07	2.13	26.70
OPTIMIS	0.06	0.91	6.17
HOIST	**0.03**	**0.29**	**3.98**

Totals include SPICE simulation and method overhead. Case-1203 values are estimated from the corresponding simulation counts and repository timing records.

Table 5: Ablation and Fusion Comparisons Across Cases: Relative Error and SPICE Cost.

Variant	Rel. Err. (%)	#SPICE calls
Full (Ours)	**8.28 / 2.59 / 4.73**	3,748 / 2,848 / 5,336
Without residual calibration	19.00 / 12.60 / 19.24	3,248 / 2,048 / 3,772
Without boundary heap	16.80 / 8.70 / 15.05	3,384 / 2,484 / 4,972
Without failure heap	23.60 / 10.90 / 18.32	3,384 / 2,484 / 4,972
Without global context	17.90 / 41.30 / 99.70	3,748 / 2,848 / 4,772
Additive fusion	11.40 / 5.90 / 8.65	3,748 / 2,848 / 5,336
Concatenation fusion	15.70 / 9.85 / 12.73	3,748 / 2,848 / 5,336

Each slash-separated entry is Case-18 / Case-243 / Case-1203. Calls include the condition-specific label ledger. The last two rows change only the fusion operator; all other settings remain fixed.

fixed-capacity heaps and reservoirs. Runtime therefore falls from 0.91 to 0.29 hours on Case-243 and from an estimated 6.17 to 3.98 hours on Case-1203. Overall, HOIST is 26%–68% faster than the fastest competing method through fewer SPICE calls and lower training and streaming-inference overhead.

5.4 Ablation Study

Table 5 shows that removing residual calibration or either retention heap consistently degrades accuracy, confirming their complementary roles. Removing global context is most damaging on Case-1203, increasing error from 4.73% to 99.7%, which highlights the importance of cross-device interaction modeling at high dimension. With all other settings fixed, multiplicative fusion outperforms additive and concatenation fusion on all three cases, supporting explicit local–global feature modulation.

5.5 Training-label Efficiency Analysis

Figure 5 shows that Case-18 saturates early, whereas the higher-dimensional cases benefit from larger initial label budgets, with Case-1203 reaching 6.3% error at 2^{11} labels. The appropriate training budget therefore increases with circuit dimension, but substantially more slowly than dimensionality itself.

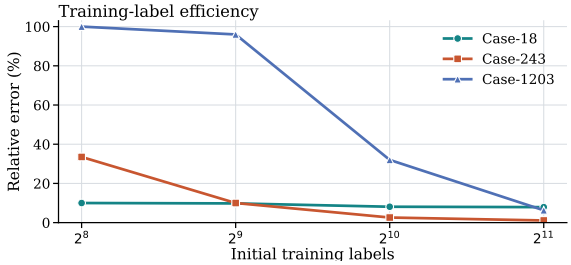

Figure 5: Training-label efficiency under fixed residual and audit budgets.

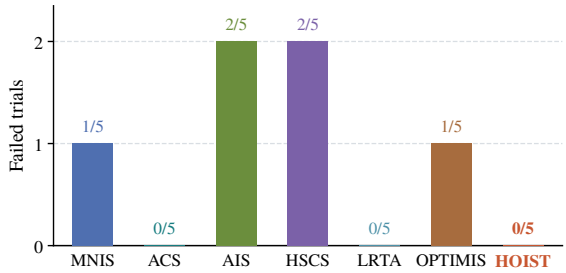

Figure 6: Case-18 robustness over five seeds; failure denotes relative error above 20%.

5.6 Robustness Study

Figure 6 summarizes the five-seed campaigns using failure counts only. HOIST, ACS, and LRTA complete all five trials successfully, whereas MNIS and OPTIMIS each fail once and AIS and HSCS each fail twice. Thus, HOIST matches the best observed repeated-seed reliability without a failed trial.

6 Conclusion

We presented HOIST, a device-structured and SPICE-validated framework for high-dimensional, high-sigma yield estimation. By combining local–global interaction modeling, complementary tail discovery, and randomized physical validation, HOIST uses surrogate predictions to guide sampling without treating them as final failure labels. Across SRAM benchmarks with 18–1203 variables, HOIST achieves 2.59%–8.28% relative error using 2,848–5,336 SPICE simulations while obtaining the lowest end-to-end runtime among the evaluated methods.

Acknowledgment

This work is done during internship in Primarius Technologies and also supported by the Start-up Research Fund of Southeast University (RF1028625045).

References

[1] M. J. Pelgrom, A. C. Duinmaijer, and A. P. Welbers, "Matching properties of mos transistors," *IEEE Journal of solid-state circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[2] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on cmos sram cell stability," *IEEE journal of Solid-state circuits*, vol. 36, no. 4, pp. 658–665, 2001.

[3] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of sram array for yield enhancement in nanoscaled cmos," *IEEE transactions on computer-aided design of integrated circuits and systems*, vol. 24, no. 12, pp. 1859–1880, 2005.

[4] Y. Liu, G. Dai, and W. W. Xing, "Seeking the yield barrier: High-dimensional sram evaluation through optimal manifold," in *2023 60th ACM/IEEE Design Automation Conference (DAC)*. IEEE, 2023, pp. 1–6.

[5] S. Yin, G. Dai, and W. W. Xing, "High-dimensional yield estimation using shrinkage deep features and maximization of integral entropy reduction," in *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, 2023, pp. 283–289.

[6] S. Shen, X. Li, Z. Liu, J. Ma, Y. Wang, Y. Wu, Y. Sun, and W. W. Xing, "Openyield: An open-source sram yield analysis and optimization benchmark suite," in *2025 IEEE 43rd International Conference on Computer Design (ICCD)*. IEEE, 2025, pp. 167–175.

[7] R. Y. Rubinstein, *Simulation and the Monte Carlo Method*. Wiley, 1981.

[8] J. A. Bucklew and J. Bucklew, *Introduction to rare event simulation*. Springer, 2004, vol. 5.

[9] S.-K. Au and J. L. Beck, "Estimation of small failure probabilities in high dimensions by subset simulation," *Probabilistic engineering mechanics*, vol. 16, no. 4, pp. 263–277, 2001.

[10] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, "Breaking the simulation barrier: Sram evaluation through norm minimization," in *2008 IEEE/ACM International Conference on Computer-Aided Design*. IEEE, 2008, pp. 322–329.

[11] X. Shi, F. Liu, J. Yang, and L. He, "A fast and robust failure analysis of memory circuits using adaptive importance sampling method," in *Proceedings of the 55th Annual Design Automation Conference*, 2018, pp. 1–6.

[12] W. Xing, Y. Liu, W. Fan, and L. He, "Every failure is a lesson: Utilizing all failure samples to deliver tuning-free efficient yield evaluation," in *Proceedings of the 61st ACM/IEEE Design Automation Conference*, 2024, pp. 1–6.

[13] Y. Liu, L. He, and W. W. Xing, "Beyond the yield barrier: Variational importance sampling yield analysis," in *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, 2024, pp. 1–9.

[14] X. Shi, H. Yan, Q. Huang, J. Zhang, L. Shi, and L. He, "Meta-model based high-dimensional yield analysis using low-rank tensor approximation," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[15] Y. Liu and W. W. Xing, "Fusis: Fusing surrogate models and importance sampling for efficient yield estimation," in *2025 Design, Automation & Test in Europe Conference (DATE)*. IEEE, 2025, pp. 1–7.

[16] X. Shi, H. Yan, J. Wang, X. Xu, F. Liu, L. Shi, and L. He, "Adaptive clustering and sampling for high-dimensional and multi-failure-region sram yield analysis," in *Proceedings of the 2019 International Symposium on Physical Design*, 2019, pp. 139–146.

[17] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," *Advances in neural information processing systems*, vol. 30, 2017.

[18] H. Hu, P. Li, and J. Z. Huang, "Enabling high-dimensional bayesian optimization for efficient failure detection of analog and mixed-signal circuits," in *Proceedings of the 56th Annual Design Automation Conference 2019*, 2019, pp. 1–6.

[19] W. G. Cochran, *Sampling techniques*. john wiley & sons, 1977.

[20] W. Wu, S. Bodapati, and L. He, "Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage," in *Proceedings of the 2016 on International Symposium on Physical Design*, 2016, pp. 153–160.