GLASS: A GLASS-BOX, LOOP-VERIFIED METHOD FOR ANALOG CIRCUIT SIZING WITH SYMBOLIC PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic analog circuit sizing is now practical, but each candidate design requires an expensive SPICE simulation, and unguided search carries little explicit design knowledge. In controlled comparisons, the failures we observe stem from what we term a cold-start trap: the agent exhausts its simulation budget in electrically degenerate regimes. We propose GLASS, a glass-box, loop-verified method that resolves this trap at initialization and preserves search momentum through stall recovery and numerical warm-starting. GLASS comprises: 1) SIZE-90K, a dataset of 90,858 sizing simulations from six amplifier topologies across four technology platforms; 2) Symbolic-Op, a calibrated behavioral symbolic layer that converts solver-derived candidates into injectable priors and audits the initial prior by re-simulation in the target testbench; and 3) Multi-Task Surrogate, a compact neural network that predicts six specification metrics and screens proposals by predicted cost. On 12 sizing tasks, GLASS reaches 60.0% strict Pass@1 and 81.7% Pass@3, surpassing the strongest evaluated baseline by 36.7 percentage points with same-topology calibration and surrogate training. Under matched proposal budgets on a single topology, the priors raise the share of search chains that meet specification from 26.3% to 56.0% and reduce the median evaluations required from 258 to 148. Across ablations, degenerate designs fall from 29/60 without symbolic guidance to 0/60 under the full operator set. We release our code at `anonymous.4open.science/r/GLASS-analog-sizing-86CF`.

1 INTRODUCTION

Analog circuit sizing determines the device dimensions of a given topology that meet target specifications. Evaluating each candidate design requires an expensive SPICE simulation, and a full sizing session can consume thousands of runs, compelling designers to rely on hand analysis to narrow the search space. The main question we address is how to provide circuit-level knowledge before the search spends its simulation budget.

Prior work on analog sizing falls into three categories. Bayesian optimization and reinforcement learning treat sizing as numerical search: they tune device parameters against simulator feedback without an internal circuit-theoretic model (Lyu et al., 2018; Ahmadzadeh et al., 2026; Zhang et al., 2022; Settaluri et al., 2020), so each new task starts from scratch.

Surrogate-assisted methods replace simulation calls with learned estimators (Budak et al., 2022; Shahane et al., 2023), but public circuit resources primarily capture netlist structure rather than the paired parameter-to-metric supervision required to train effective sizing surrogates (Li et al., 2025; Tao et al., 2024; Pan et al., 2026b; Bhandari et al., 2025; Gao et al., 2025). Symbolic and equation-based methods derive interpretable design equations from circuit theory, calibrate them against simulation, and use them to guide sizing (Chen et al., 2026; Ghosh et al., 2026; Bujana & Karsilayan, 2026; Abel et al., 2022) (see Appendix A for a detailed discussion).

Recent agentic methods belong to the same black-box category: they deploy language models to parse netlists, propose parameter updates, and query simulators in a closed loop (Pan et al., 2026a; Ngo et al., 2026; Shen et al., 2026; Liu et al., 2024; Wu et al., 2025). Nevertheless, two bottlenecks persist: 1) Without circuit-theoretic guidance, the loop readily falls into a cold-start trap: the agent

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.

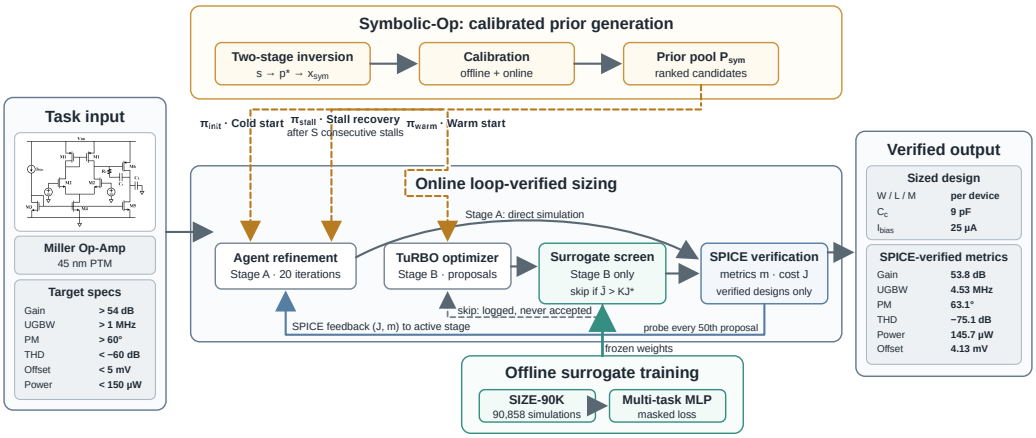


Figure 1: The **GLASS** pipeline maps a sizing task to a SPICE-verified design. Calibrated priors guide the search, and a frozen multi-task surrogate screens proposals. **Top**: **Symbolic-Op** derives a calibrated prior pool $\mathcal{P}_{\text{sym}}$. **Bottom**: **SIZE-90K** trains the surrogate through masked multi-task learning. **Center**: Stage A simulates every agent design, while Stage B screens TuRBO proposals and periodically probes SPICE. The core design combines three symbolic injection points with surrogate screening, but accepts a candidate only after SPICE verification.

proposes dimensions that push transistors out of active saturation into cut-off or triode regimes, where the amplifier cannot function. In these regimes, SPICE returns no usable measurement, and the search spends its budget without progress. 2) The intermediate reasoning remains heuristic and informal, leaving the rationale behind compliant solutions unverifiable.

Symbolic methods address this verifiability bottleneck by construction, as their design equations explicitly expose the rationale behind each proposal, but they stop at generating analytical solutions. Existing approaches do not inject analytical priors as pluggable operators at distinct timings such as initialization, stall recovery, and numerical warm start. They also do not audit whether a prior from one environment remains valid on a distinct target testbench. GLASS occupies this neuro-symbolic middle ground: rather than relying solely on unguided language models or static symbolic solvers, it injects calibrated analytical priors into the agent loop at three controlled intervention points, and it accepts no design without transistor-level SPICE verification.

To overcome these challenges:

1. We propose **GLASS**, a glass-box, loop-verified method for analog circuit sizing with symbolic priors, in which a calibrated symbolic layer resolves the cold-start trap at initialization and preserves search momentum through stall recovery and numerical warm-starting.

2. We construct **SIZE-90K**, a dataset of 90,858 transistor-level sizing simulations from six amplifier topologies across four technology platforms, each pairing a device-dimension vector with its simulated metrics.

3. We develop **Symbolic-Op**, a calibrated behavioral symbolic layer that converts solver-derived candidates into injectable priors, with gain and UGBW calibration, online correction, and a seed-transfer audit that re-simulates the initial prior.

4. We integrate a **Multi-Task Surrogate** that predicts the six specification metrics from device dimensions and screens proposals to save simulator budget, together with a loop-verification scheme: the surrogate screens proposals by estimated cost, and three safeguards bound the verification rate so the loop accepts no design without re-simulation.

We present GLASS in Fig. 1 and evaluate it on 12 sizing tasks spanning six topologies under one operator set and one shared surrogate. Each task runs five trials under a fixed proposal cap. GLASS reaches 60.0% strict Pass@1 and 81.7% Pass@3, surpassing the strongest baseline by 36.7 percentage points with same-topology calibration and surrogate training.

2 PROBLEM FORMULATION

Consider the sizing problem for a fixed analog circuit topology. Let $\mathbf{x} \in \mathcal{X}$ denote the vector of continuous and discrete design parameters, and let $\mathbf{m}(\mathbf{x})$ denote the performance metrics measured from the resulting circuit. We define a sizing task by a specification target $\mathbf{s}$ over six standard analog performance metrics: DC gain (dB), unity-gain bandwidth (UGBW, MHz), phase margin (PM, °), total harmonic distortion (THD, dB), input-referred offset (mV), and power consumption (mW). We measure each candidate with Ngspice and treat the cumulative number of simulation calls as the principal resource constraint in this study.

In an agentic sizing loop, the agent proposes an updated parameter vector $\mathbf{x}$ at each iteration. The environment then instantiates it into a netlist, runs the transistor-level simulation, and returns the measured metrics $\mathbf{m}(\mathbf{x})$ as feedback to guide future proposals, so sample efficiency depends on how informative each simulation call is.

To objectively rank candidate designs across heterogeneous specifications and their distinct native units, we define a scalar reference cost $J(\mathbf{x})$:

$$J(\mathbf{x}) = \frac{P(\mathbf{x})}{P_{\max}} + \sum_{k \in \mathcal{K}} \lambda_k \max\big(0,\ \mathrm{viol}_k(\mathbf{x})\big) + \eta(\mathbf{x}), \tag{1}$$

where $P(\mathbf{x})$ denotes the dissipated power, $P_{\max}$ denotes the specified upper power limit, and $\mathcal{K}$ indexes the constrained metrics. The hinge term $\mathrm{viol}_k(\mathbf{x})$ measures the shortfall of metric k in its native unit, $\max(0, s_k - m_k(\mathbf{x}))$ for higher-is-better metrics and $\max(0, m_k(\mathbf{x}) - s_k)$ otherwise, and λ_k absorbs the scale of that unit. We size the weights so that any substantive violation pushes the cost above one, as the power term alone stays below one. Offset labels are the sparsest, so the offset weight is largest (see Table A2). The barrier penalty $\eta(\mathbf{x})$ adds 100 when DC gain or UGBW falls below 10 dB and 0.1 MHz, penalizing low-gain or low-bandwidth designs.

Every term in Equation (1) is non-negative, so J penalizes violations but does not certify compliance. The loop ranks candidates by this cost and uses a strategy-weighted variant as its objective, while the tolerance-adjusted pass rule of Section 4 decides compliance.

3 METHOD: GLASS

GLASS augments the agentic sizing loop through two core components. First, Symbolic-Op converts analytical circuit formulations into a calibrated pool of domain priors and injects them at three injection points: cold-start initialization, stall recovery during refinement, and warm-starting the downstream numerical optimizer. Second, trained on our curated SIZE-90K dataset, a Multi-Task Surrogate predicts the six specification metrics of a candidate design from its device dimensions and screens proposals by their estimated reference cost, ensuring that unviable proposals do not consume expensive SPICE evaluations (see Algorithm 1).

3.1 SIZE-90K

Trajectory-Grounded Dataset Construction. We curate SIZE-90K from closed-loop search trajectories, aggregating 90,858 transistor-level simulations from six topologies across four technology platforms (see Table 1): the SkyWater SKY130 process design kit (PDK), 45 nm and 180 nm predictive technology models (PTM), and discrete MOSFET models. The offline budget does not enter our comparisons, which count only search-time simulations. Each chain saves 74–110 evaluations, so the offline budget breaks even only after 800–1,200 chains.

Multi-Task Profiles and Sparse Labels. Large-signal transient (THD) and statistical operating-point (offset) evaluations are far more expensive than small-signal AC analyses. SIZE-90K preserves this hierarchy, so the expensive metrics carry sparse labels, which motivates the masked multi-task loss of Section 3.4. Simulations that fail to converge record a UGBW sentinel of 10^9 or above, and we treat the 1,365 such rows as missing. The masked multi-task loss of Section 3.4 ignores these rows, so the sentinel never enters training or the fidelity counts.

3.2 SYMBOLIC-OP

Candidate Generation via Two-Stage Inversion. Symbolic-Op generates the prior pool $\mathcal{P}_{\text{sym}}$ from a specification $\mathbf{s}$ in two stages. Instead of solving directly for device dimensions, it parametrizes the circuit by its small-signal operating parameters $\mathbf{p}$, such as stage transconductances, output conductances, and compensation capacitances, and a behavioral model $\mathbf{g}(\mathbf{p})$ maps those parameters to the metrics. The first stage inverts the model with a constrained solve, $\mathbf{p}^* = \arg\min_{\mathbf{p}} \|\mathbf{g}(\mathbf{p}) - \mathbf{s}\|_2^2$, subject to the stability constraints of the topology. The second stage maps the solution to transistor dimensions through g_m/I_D lookup tables, $\mathcal{T}(\mathbf{p}^*) = \mathbf{x} = (W, L, M)$. Since the solver is deterministic per target, we sweep the bandwidth target over scaling factors and populate the pool with candidates trading speed against power.

Simulation-Grounded Calibration. Analytical formulations deviate from nanometer-scale SPICE physics, so Symbolic-Op calibrates the behavioral model against simulation. For metric k, we apply a two-stage scalar calibration $\hat{m}_k = f_k(\tilde{m}_k) + \delta_k$, where $\tilde{m}_k$ denotes the raw behavioral prediction. We fit the baseline mapping function f_k offline using least squares, employing linear regression for gain, logarithmic regression for UGBW, and a constant offset for PM. The coefficients come from the training trajectories that form SIZE-90K.

During closed-loop execution, we correct the calibration online by the median residual over past simulations, setting $\delta_k = \text{median}_t\{m_k^{\text{spice}}(\mathbf{x}_t) - f_k(\tilde{m}_k(\mathbf{x}_t))\}$, with UGBW updated in ratio form, where the running median guards against transient outliers. Bias-current and device-size assignments come directly from the analytical solver and we do not calibrate them, a gap that the seed-transfer audit exposes on one topology (see Section 4.5).

Across the 12 tasks, offline calibration reduces the median gain deviation by a factor of 3.9, from 10.3 to 2.7 dB, and increases the share of gain predictions within 3 dB from 26% to 52%. Even after calibration, the model errs substantially on the remaining metrics: median UGBW and PM errors are 54.0% and 52.5°, against 56.0% and 60.8° without calibration. The loop therefore verifies every adopted design with SPICE rather than trusting the symbolic prediction.

Seed-Transfer Audit. Symbolic-Op incorporates an empirical seed-transfer audit to validate injected priors under realistic operating conditions. Specifically, the audit instantiates the initial candidate's parameter vector directly into the target verification AC testbench and re-simulates it. This cross-check effectively decouples initialization defects from failures in downstream searches. If a prior's simulated performance collapses under the target testbench, we identify it as an unviable seed, turning a silent failure into an attributed one. The audit is diagnostic rather than corrective: it does not, on its own, replace a defective prior.

3.3 INJECTION POINTS

We formalize the three interventions as operators π_{init}, π_{stall}, and π_{warm}, each of which draws from the symbolic pool $\mathcal{P}_{\text{sym}}$ under a stage-specific rule and modifies the candidate. Let $\text{GLASS}(S)$ denote the same loop restricted to an operator subset $S \subseteq \{\pi_{\text{init}}, \pi_{\text{stall}}, \pi_{\text{warm}}\}$. No-symbolic and seed-only are $\text{GLASS}(\emptyset)$ and $\text{GLASS}(\{\pi_{\text{init}}\})$, respectively.

Initialization (π_{init}). The initial prior replaces the agent's unconditioned configuration via masked substitution, $\text{MERGE}(\mathbf{x}_{\text{sym}}, \mathbf{x}_{\text{agent}}) = \mathbf{M} \odot \mathbf{x}_{\text{sym}} + (\mathbf{I} - \mathbf{M}) \odot \mathbf{x}_{\text{agent}}$, where $\mathbf{M} \in \{0,1\}^{|\mathcal{X}|}$ denotes the support indicator of parameters specified by the symbolic prior. Unspecified parameters keep their agent values, so the netlist stays complete. If candidate generation fails or yields an infeasible design, the system defaults to the agent's seed.

Stall Recovery (π_{stall}). During the refinement phase, the agent retains primary control over the optimization trajectory, and symbolic priors intervene only when the search stagnates. Stall recovery is event-triggered, not periodic: injection fires only after S consecutive iterations without reference cost improvement. Any cost reduction immediately resets the counter and defers injection, so an injection never overrides a productive trajectory. An early variant used a fixed-period injection schedule, which caused disruptive oscillations between agent-driven exploration and symbolic overrides. Conditioning the injection on stagnation prevents this interference and preserves the agent's local search momentum.

Table 1: Summary of **SIZE-90K**. Coverage is the share of simulation rows with a valid label.

Item	Value
Simulations	90,858
Topologies $\times$ platforms	6×4
Gain, PM	99.8%
UGBW	98.3%
THD	55.3%
Power	41.4%
Offset	4.3%

Table 2: The three symbolic injection operators of GLASS, their triggers, and the quantity each one injects. Injected payloads merge into the current candidate.

Operator	Trigger	Injected payload
π_{init}	Before the first simulation	Nominal symbolic candidate with masked merge
π_{stall}	S stalled iterations	Next Pareto candidate from the symbolic pool
π_{warm}	At handover to optimization	Candidate with the highest verified gain

Warm Start (π_{warm}). When transitioning from agentic search to numerical optimization, we merge the highest-gain simulated candidate into the initialization design to seed the TuRBO optimizer (Eriksson et al., 2019), as gain drives the cold-start collapses.

All three intervention stages draw from a unified symbolic pool $\mathcal{P}_{\text{sym}}$ (see Table 2 for their triggers and payloads). The seed-transfer audit probes the margin condition of Proposition 1 (see Section 4.5), so a broken mapping becomes visible rather than silent.

3.4 Multi-Task Surrogate

Input Representation and Masked Multi-Task Prediction. We represent each candidate design as a standardized parameter vector $\mathbf{x}$ comprising transistor gate widths, lengths, multiplicities, and passive component values. The surrogate employs a shared multi-layer perceptron (MLP) backbone coupled with seven independent linear heads $\hat{f}_k$ to concurrently predict the target metrics alongside an auxiliary DC output voltage. A dedicated AC refinement head additionally estimates gain, UGBW, and PM. To accommodate heterogeneous evaluation coverage in SIZE-90K without synthetic imputation, we train the model via an element-wise masked mean squared error (MSE) plus an auxiliary AC loss on the refinement head for the three AC metrics:

$$\mathcal{L}(\Theta) = \sum_{k=1}^{7} \frac{\sum_{i=1}^{B} M_{i,k} \big(\hat{y}_{i,k} - z_k(y_{i,k})\big)^2}{\sum_{i=1}^{B} M_{i,k}} + \frac{1}{2}\mathcal{L}_{\text{ac}}, \tag{2}$$

where $M_{i,k} \in \{0, 1\}$ indicates whether SIZE-90K records metric k for sample i, and $z_k(\cdot)$ standardizes the targets, applying logarithmic warping to UGBW, THD, power, and offset. The auxiliary term $\mathcal{L}_{\text{ac}}$ computes the MSE of the AC refinement head over samples with valid AC simulations. The AC refinement head adds dedicated capacity for densely sampled AC metrics without conflicting objectives, while the shared trunk maintains multi-task generalization (see Appendix E for the layer dimensions and training hyperparameters).

We evaluate predictive fidelity with 5-fold cross-validation over the training rows, with an internal partition for early stopping. As reported in Appendix E (see Table A3), the surrogate reaches its highest raw-scale fidelity on DC gain ($R^2 = 0.737$). Target warping is indispensable for heavy-tailed metrics (UGBW, power, THD, offset): raw-scale fits are degenerate ($R^2 < 0$) under outlier sensitivity, while log-space models reach $R^2 = 0.847$ on THD and 0.757 on offset. In Appendix E.1 we present a session-grouped variant that keeps all rows of one session on the same side of the split: gain R^2 drops from 0.737 to 0.285 and PM from 0.672 to 0.193. We therefore treat the surrogate as an in-distribution ranking filter, not a cross-session predictor.

Candidate Screening and Verification Safeguards. During closed-loop optimization, the surrogate serves as an early-stage gatekeeper, pruning unviable candidates before expensive SPICE invocations. For any proposal $\mathbf{x}$, the surrogate computes an estimated reference cost $\hat{J}(\mathbf{x})$ via Equation (1). Candidates with $\hat{J}(\mathbf{x}) > K \cdot J^*$ bypass simulation, and the loop logs them as skipped, so they consume no simulator calls.

Three safeguards prevent approximation errors from derailing the trajectory:

- **Ground-Truth Anchoring**: The reference baseline J^* updates strictly upon verified SPICE simulations. Surrogate predictions can therefore never autonomously tighten or corrupt the filtering threshold.
- **Simulation-Bounded Skips**: Cumulative skipped evaluations cannot exceed verified simulations (skips $< N_{\text{eval}}$), so the loop balances every skip with a verified simulation and the optimizer cannot drift into prolonged ungrounded exploration.
- **Periodic Probing**: We route candidate proposals to SPICE simulation at a fixed interval T_{probe}, regardless of their predicted cost $\hat{J}(\mathbf{x})$.

We accept no skipped proposal as a compliant solution, and we accept no design without a transistor-level simulation. During refinement, the loop simulates every committed design, including each stall-recovery merge, at the start of the next iteration. After t proposals, GLASS therefore executes at least $\lfloor t/2 \rfloor$ physical simulations with at least $\lfloor t/T_{\text{probe}} \rfloor$ deterministic probes, ruling out arbitrarily long unverified stretches (see Appendix D for the loop constants).

Algorithm 1 The GLASS sizing loop.

Require: Specification $\mathbf{s}$, initial design $\mathbf{x}_{\text{init}}$, symbolic pool $\mathcal{P}_{\text{sym}}$, surrogate $\hat{J}$, constants $(S, K, T_{\text{probe}}, N_{\text{agent}}, N_{\text{turbo}})$

$\mathbf{x} \leftarrow \textsc{Merge}(\mathcal{P}_{\text{sym}}[0], \mathbf{x}_{\text{init}})$ ▷ Cold-start prior (π_{init})
$(J^*, \mathbf{x}^*, \text{stalls}, \text{skips}, N_{\text{eval}}, i) \leftarrow (\infty, \perp, 0, 0, 0, 1)$
for $t = 1$ **to** N_{agent} **do** ▷ Stage A: agent refinement
  $J \leftarrow \textsc{Simulate}(\mathbf{x})$; $N_{\text{eval}} \leftarrow N_{\text{eval}} + 1$
  $(J^*, \mathbf{x}^*, \text{stalls}) \leftarrow (J, \mathbf{x}, 0)$ **if** $J < J^*$ **else** $(J^*, \mathbf{x}^*, \text{stalls} + 1)$
  if $\mathbf{x}$ passes the specification **then**
    break
  end if
  $\mathbf{x} \leftarrow \textsc{AgentProposal}(\mathbf{x})$
  if stalls $\geq S$ **then** ▷ Stall recovery (π_{stall})
    $\mathbf{x} \leftarrow \textsc{Merge}(\mathcal{P}_{\text{sym}}[i \bmod |\mathcal{P}_{\text{sym}}|], \mathbf{x})$; $i \leftarrow i + 1$; stalls $\leftarrow 0$
  end if
end for
$\mathbf{x}_0 \leftarrow \textsc{Merge}(\arg\max_{\mathbf{c} \in \mathcal{P}_{\text{sym}}} \text{gain}(\mathbf{c}), \mathbf{x}_{\text{init}})$; seed the optimizer with $\mathbf{x}_0$ and the three best verified Stage-A designs ▷ Warm start (π_{warm})
for $u = 1$ **to** N_{turbo} **do** ▷ Stage B: numerical optimization
  $\mathbf{x}' \leftarrow \textsc{TuRBOProposal}()$
  if screening **and** $\hat{J}(\mathbf{x}') > K \cdot J^*$ **and** skips $< N_{\text{eval}}$ **and** $u \bmod T_{\text{probe}} \neq 0$ **then**
    skips $\leftarrow$ skips $+ 1$ ▷ Record predicted cost only
  else
    $J \leftarrow \textsc{Simulate}(\mathbf{x}')$; $N_{\text{eval}} \leftarrow N_{\text{eval}} + 1$
    $(J^*, \mathbf{x}^*) \leftarrow (J, \mathbf{x}')$ **if** $J < J^*$ **else** $(J^*, \mathbf{x}^*)$
  end if
end for
return $\mathbf{x}^*$ ▷ Best verified design

4 Experiments

4.1 Experimental Settings

Following Pan et al. (2026a), we evaluate GLASS on 12 sizing tasks: a two-stage Miller operational amplifier (Miller Op-Amp) in a 45 nm PTM (Ahmadzadeh et al., 2026), a two-stage discrete-MOSFET operational amplifier (Discrete Op-Amp), a complementary-input class-AB operational amplifier (Class-AB Op-Amp) in a 180 nm PTM (Liu & Chitnis, 2026), a damping-factor-control frequency-compensation amplifier (DFCFC) (Li et al., 2025; Leung & Mok, 2001), a nested Miller-compensated amplifier with nulling resistor (NMCNR) (Li et al., 2025; Leung & Mok, 2001), and a positive-feedback-compensated amplifier (PFC) (Li et al., 2025; Ramos & Steyaert, 2002). DFCFC, NMCNR, and PFC use the SkyWater SKY130 PDK. Each circuit carries two specification sets, and every task runs five trials (see Table A1).

Table 3: Strict $Pass@1$ and $Pass@3$ (%) over 12 sizing tasks ($n = 5$ runs per task). **Bold** denotes the best result in each row. The pooled $Pass@1$ row summarizes all 60 trials per method, and pooled $Pass@3$ macro-averages the 12 task-level values, with 95% task-level cluster bootstrap intervals.

			GLASS		**VLM-CAD**		**EEsizer**	
Circuit	Tech.	Spec	P@1	P@3	P@1	P@3	P@1	P@3
Miller Op-Amp	PTM 45 nm	SpecsA	**100**	**100**	0	0	0	0
		SpecsB	**40**	**90**	0	0	0	0
Discrete Op-Amp	Discrete MOSFET	SpecsA	**60**	**100**	40	90	0	0
		SpecsB	**40**	**90**	**40**	**90**	0	0
Class-AB Op-Amp	PTM 180 nm	SpecsA	**60**	**100**	**60**	**100**	20	60
		SpecsB	**100**	**100**	40	90	20	60
DFCFC	SKY130	SpecsA	**80**	**100**	0	0	0	0
		SpecsB	**80**	**100**	60	**100**	0	0
NMCNR	SKY130	SpecsA	**60**	**100**	0	0	0	0
		SpecsB	**100**	**100**	0	0	0	0
PFC	SKY130	SpecsA	0	0	**40**	**90**	0	0
		SpecsB	0	0	0	0	0	0
Pooled			**60.0**	**81.7**	23.3	46.7	3.3	10.0
95% CI			[40.0, 78.3]	[57.5, 99.2]	[10.0, 36.7]	[22.5, 71.7]	[0.0, 8.3]	[0.0, 25.0]

We adopt $Pass@k$, the probability that at least one correct solution appears among k generations that we sample independently for a task:

$$Pass@k = 1 - \frac{C_{n-c}^{k}}{C_{n}^{k}}, \tag{3}$$

where n denotes the total number of attempts, c the number of correct generations, and C_n^k the binomial coefficient. We set $n = 5$, and we report pooled $Pass@3$ as the macro-average of the 12 task-level $Pass@3$ values. We report 95% confidence intervals (CIs) from a task-level cluster bootstrap over the 12 tasks. A design passes a task when every metric meets its target within a 5% one-sided tolerance, and we refer to this tolerance-adjusted criterion as strict Pass@1. For a fair comparison, we use EEsizer (Liu & Chitnis, 2026) and VLM-CAD (Pan et al., 2026a) as baselines, and all methods use Gemma-4-31B-it (Team et al., 2026) as their backbone language model, and we run every circuit simulation in Ngspice.

4.2 Main Results

Benchmark Sizing Performance. We report strict $Pass@1$ and $Pass@3$ against external baselines on the 12 tasks (see Table 3). GLASS achieves a pooled strict $Pass@1$ of 60.0%, outperforming VLM-CAD (23.3%) by 36.7 percentage points and EEsizer (3.3%) by 56.7 percentage points. GLASS exceeds the best baseline on 8 tasks and ties it on 3. GLASS also outperforms the ablation controls, the no-symbolic arm (10.0%, Fisher exact test $p = 9.2 \times 10^{-9}$) and the seed-only variant (31.7%, $p = 0.0032$), showing that closed-loop injection contributes margin beyond initialization alone (see Section 4.3). A sign test over task clusters gives $p = 0.039$ against VLM-CAD and $p = 0.002$ against EEsizer.

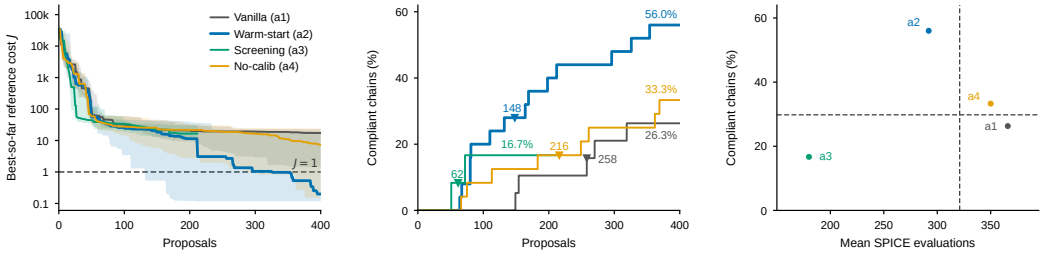
Search Sample Efficiency. Beyond single-trial success, pooled $Pass@3$ reaches 81.7%, indicating that three trials suffice for most tasks. Each trial has the same 400-proposal cap, and we estimate $Pass@3$ from the five completed trials. One source of remaining failures is PFC, which we diagnose via the seed-transfer audit of Section 4.5.

4.3 Ablation

Controlled Ablation Arms. We evaluate three configurations of the operator set defined in Section 3.3 (see Table 4), where comparing seed-only with GLASS isolates the marginal gain of closed-loop interventions. We benchmark the three configurations on all 12 sizing tasks (five independent trials per task) under an identical agentic pipeline, language model backbone, proposal cap, agent scaffold, and prompts, with no component tuned per arm:

Table 4: Ablation over symbolic injection operators ($n = 5$ trials per task, 60 runs per arm). Checkmarks mark the enabled operators. Degenerate denotes runs that report non-positive DC gain (≤ 0 dB) or fail the measurement. In our records every degenerate run fails the measurement, so the trap concerns unmeasurable designs. **Bold** denotes the best value per column.

Arm	init	stall	warm	P@1	P@3	Degenerate
GLASS	✓	✓	✓	**60.0**	**81.7**	**0/60**
seed-only	✓	×	×	31.7	50.8	7/60
no-symbolic	×	×	×	10.0	27.5	29/60



(a) Median best-so-far reference cost J across independent chains (band: interquartile range). The dashed line marks the $J = 1$ reference-cost guide. We carry final costs forward after early exit.

(b) Cumulative percentage of search chains reaching specification compliance. Downward triangles indicate and label the median proposals required to achieve first compliance.

(c) Mean simulator calls against chain success rate. Dashed guides mark the median of the four configurations on each axis. The upper-left quadrant favors fewer simulator calls and a higher success rate.

Figure 2: Matched-budget search compression on DFCFC SpecsA under a fixed cap of 400 proposals per chain. We score compliance strictly via SPICE.

- **GLASS** ($\{\pi_{\text{init}}, \pi_{\text{stall}}, \pi_{\text{warm}}\}$): Enables cold-start initialization, stall recovery during search stagnation, and warm-starting the downstream numerical optimizer.
- **seed-only** ($\{\pi_{\text{init}}\}$): Injects the calibrated symbolic prior strictly at cold start, disabling subsequent closed-loop interventions.
- **no-symbolic** ($\emptyset$): Removes the symbolic layer entirely, leaving the loop without guidance while still operating under the full tool protocol within the same proposal cap.

Symbolic Priors as Operating-Regime Regularizers. Comparing no-symbolic with seed-only exposes the cold-start trap: without initialization priors, the unconstrained agent collapses into electrically degenerate designs in nearly half the runs (29/60). Initializing with π_{init} anchors the search inside an active saturation basin and reduces degeneracy to 7/60, while closed-loop injection prevents unrecoverable collapses entirely (0/60), and zero-tolerance re-scoring preserves the ordering (see Appendix C), so the trend is monotone in the operator count.

Failure Diagnosis and Residual Margins. Since GLASS produced no degenerate failures, all non-passing solutions avoid the degeneracy criterion of Table 4. Under a diagnostic relaxed criterion that permits a single non-power metric shortfall within tolerance, GLASS's success rate rises from 60.0% to 83.3%. This 23.3-percentage-point increase indicates that the relaxed criterion rescues 14 of the 24 remaining strict failures.

Online Operators and Topological Boundaries. The closed-loop operators (π_{stall}, π_{warm}) add 28.3 percentage points in strict Pass@1 (31.7% to 60.0%) and 30.9 points in Pass@3 (50.8% to 81.7%). On the Discrete Op-Amp, GLASS and seed-only tie at 50% against 0% for no-symbolic. In this discrete-component setting, the cold-start prior provides the decisive boost, while the closed-loop operators add no measurable margin (50% vs 50%, five trials).

4.4 Matched-Budget Compression

We benchmark individual GLASS components on DFCFC SpecsA under an equal cap of 400 numerical proposals per chain (see Fig. 2 and Table A5). Symbolic warm-starting (a2) more than doubles the chain success rate of vanilla search (a1), from 26.3% to 56.0%, while cutting median evaluations to compliance from 258 to 148, a reduction of roughly 43%.

Calibration drives a large share of this acceleration: removing it (a4) incurs a 22.7-percentage-point drop in success rate (to 33.3%), consistent with uncalibrated behavioral models injecting distorted initializations that impede downstream convergence. Meanwhile, surrogate screening (a3) compresses simulator demand to 180 evaluations per chain. Its lower success rate (16.7%) aligns with its function as a conservative pruning filter within local trajectories rather than a generative source of new designs (see Appendix E.1). Under the same budget, no configuration reaches specification on the more complex NMCNR and PFC topologies (see Appendix G), delineating the operating boundary of single-chain budget compression (see Fig. 2c).

4.5 Seed-Transfer Audit

We execute a post-hoc seed-transfer audit to diagnose the PFC failures. Proposition 1 (see Appendix J) gives the condition: when the calibrated model predicts every metric within an error bound and the seed's margins exceed that bound, the seed still meets every target under simulation. On NMCNR and DFCFC the prior transfers reliably (15/15 runs), while on PFC it predicts 84.2 dB but collapses to -164.8 dB, isolating a static mapping defect. We retain PFC in all pooled benchmark figures (60.0% strict Pass@1). Excluding it yields 72.0% (36/50) across the remaining 10 tasks (see Appendix H for the audited runs and the controls).

4.6 Where the gain comes from

Warm start (a2) reaches first compliance in a median of 148 evaluations against 258 for vanilla (a1), records fewer stall events per chain (49 against 68), and improves the best-so-far within 10 evaluations in about half of its stall episodes. Both patterns are consistent with the warm-start seed starting refinement from a simulated, verified candidate rather than a cold start.

5 Discussion

Why the loop holds up. The symbolic layer acts primarily as an operating-regime regularizer rather than a local fine-tuner: degenerate designs drop from 29/60 under unguided search to 7/60 with a cold-start prior, and to 0/60 under full closed-loop injection. When an analytical prior fails, a single verification simulation exposes the error rather than hiding it (see Section 4.5). Since every candidate advanced by the loop must undergo transistor-level SPICE re-simulation, analytical mapping discrepancies and surrogate approximation errors degrade only search efficiency, never the validity of adopted designs under the stated device models and testbenches.

Limitations and future work. First, our pipeline lacks an automated pre-injection gatekeeper: a pre-deployment simulation could intercept mapping failures such as the PFC bias collapse, but we run this audit as a post-hoc diagnostic. Second, our matched-budget compression remains demonstrated on a single topology, as exploratory trials across additional circuits did not reproduce equivalent efficiency gains (see Appendix G). Third, the multi-task surrogate acts as an in-distribution filter, and its predictive fidelity degrades on unseen sessions, which motivates our simulation-bounded safeguards (see Appendix E.1). Finally, Symbolic-Op needs a hand-derived behavioral model per topology, and a wrong mapping can collapse the seed (see Section 4.5).

6 Conclusion

We present GLASS, a glass-box, loop-verified method for analog circuit sizing with symbolic priors. A multi-task surrogate screens unpromising proposals under simulation-bounded safeguards, so every adopted design remains SPICE-verified. Across 12 benchmark sizing tasks, GLASS achieves 60.0% strict Pass@1 and 81.7% Pass@3, outperforming the strongest evaluated baseline by 36.7 percentage points with same-topology calibration and surrogate training. Under matched proposal budgets on a single topology, symbolic guidance raises compliant chains from 26.3% to 56.0% and cuts median evaluations from 258 to 148. When analytical formulations diverge, the post-hoc seed-transfer audit surfaces the mismatch rather than hiding it.

AI USE STATEMENT

In this work, DeepSeek-V4.1-Flash (DeepSeek-AI et al., 2026) assists in implementing experimental pipelines, analysis scripts, and simulation log processing, and in interpreting aggregate results. Gemini 3.8 Flash (Google, 2026) helps retrieve related literature, organize bibliographic citations, and draft and refine portions of the manuscript and figures. We do not use generative AI tools for empirical measurements: we execute every SPICE simulation and analytical solve, re-running the full evaluation protocol after every code modification. The agent inside the sizing loop is itself an object of study, not an assistant we use for this analysis. We review all AI-assisted outputs against our own experimental records and assume full responsibility for the final contents of this submission, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

We link to an anonymous code repository in the abstract. The repository contains the Symbolic-Op models and operators, the surrogate model and checkpoint, and the screening rule. It also contains calibration tables for two topologies, the seed-transfer audit note, and runnable demos. We list the target specifications across all 12 benchmark sizing tasks (see Table A1). Appendix D lists the closed-loop constants and symbolic scaling rules, while Appendix E details the surrogate neural architecture and training recipes. Section 4 outlines the baseline search protocol, which includes equal proposal caps, five independent trials per task, and fixed random seeds. Section 4.4 establishes the matched-budget compression framework.

Section 3.4 and Appendix E.1 describe the surrogate validation protocol and session-grouped sensitivity analyses. Since the symbolic formulation is deterministic for any target specification, the injected priors follow from per-topology behavioral models built with the procedure in Section 3.2. Finally, a single NVIDIA RTX 4060 Laptop GPU trains the surrogate, while two AMD EPYC 75F3 workstations (each assigned 16 cores) execute all SPICE simulations.

ETHICS STATEMENT

This work investigates automatic analog circuit sizing using publicly available technologies and netlists. The study engages no human participants, collects no personal data, and accesses no proprietary artifacts. Accordingly, the methodology presents no specific ethical risks or dual-use concerns, as its primary objective is to reduce simulation overhead in standard engineering optimization workflows that use publicly available process design kits and device models.

REFERENCES

Inga Abel, Maximilian Neuner, and Helmut E. Graeb. A hierarchical performance equation library for basic op-amp design. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(7):1976–1989, 2022. doi: 10.1109/TCAD.2021.3101691.

Mohsen Ahmadzadeh, Jan Lappas, Norbert Wehn, and Georges Gielen. Anacraft: Duel-play probabilistic-model-based reinforcement learning for sample-efficient pvt-robust analog circuit sizing optimization. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 45(2):901–914, 2026. doi: 10.1109/TCAD.2025.3582175.

Jitendra Bhandari, Vineet Bhat, Yuheng He, Hamed Rahmani, Siddharth Garg, and Ramesh Karri. Masala-chai: A large-scale spice netlist dataset for analog circuits by harnessing ai, 2025. URL https://arxiv.org/abs/2411.14299.

Ahmet Faruk Budak, Miguel Gandara, Wei Shi, David Z. Pan, Nan Sun, and Bo Liu. An efficient analog circuit sizing method based on machine learning assisted global optimization. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(5):1209–1221, 2022. doi: 10.1109/TCAD.2021.3081405.

Antonio J. Bujana and Aydin I. Karsilayan. A self-calibrating framework for analog circuit sizing using llm-derived analytical equations, 2026. URL https://arxiv.org/abs/2604.07387.

Yu Cao. *Predictive technology model for robust nanoelectronic design*. Springer Science & Business Media, 2011.

Zihao Chen, Ziyi Sun, Jiayin Wang, Ji Zhuang, Jinyi Shen, Xiaoyue Ke, Li Shang, Xuan Zeng, and Fan Yang. Llm-assisted op-amp behavioral-level design via agentic human-mimicking reasoning, 2026. URL https://arxiv.org/abs/2601.21321.

DeepSeek-AI et al. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

R Timothy Edwards. Google/skywater and the promise of the open pdk. In *Workshop on Open-Source EDA Technology*, volume 124, pp. 132, 2020.

David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6c990b7aca7bc7058f5e98ea909e924b-Paper.pdf.

Yu Feng, Nathaniel Weir, Kaj Bostrom, Sam Bayless, Darion Cassel, Sapana Chaudhary, Benjamin Kiesl-Reiter, and Huzefa Rangwala. Vericot: Neuro-symbolic chain-of-thought validation via logical consistency checks. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 133170–133208, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d7b3cef7c31b94a4a533db83d01a8882-Paper-Conference.pdf.

Jian Gao, Weidong Cao, Junyi Yang, and Xuan Zhang. Analoggenie: A generative engine for automatic discovery of analog circuit topologies. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 96496–96514, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/f0552f14388d95b19740dee809f5cad1-Paper-Conference.pdf.

Subhadip Ghosh, Ramesh Harjani, and Sachin S. Sapatnekar. Nemesis: Netlist-driven modeling and equation synthesis with inversion-aware spice anchoring, 2026. URL https://arxiv.org/abs/2607.05657.

Google. Introducing gemini 3.8 flash and 3.8 flash cyber, 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/3-8-flash-and-3-8-flash-cyber/.

Nan Jiang, Ziyi Wang, and Yexiang Xue. Egg-sr: Embedding symbolic equivalence into symbolic regression via equality graph. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 13650–13689, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1704fe7aaff33a54802b83a016050ab8-Paper-Conference.pdf.

Ka Nang Leung and P.K.T. Mok. Analysis of multistage amplifier-frequency compensation. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 48(9):1041–1056, 2001. doi: 10.1109/81.948432.

Jintao Li, Haochang Zhi, Ruiyu Lyu, Wangzhen Li, Zhaori Bi, Keren Zhu, Yanhan Zeng, Weiwei Shan, Changhao Yan, Fan Yang, Yun Li, and Xuan Zeng. Analoggym: An open and practical testing suite for analog circuit synthesis. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, ICCAD '24, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400710773. doi: 10.1145/3676536.3697117. URL https://doi.org/10.1145/3676536.3697117.

Qian Li, Yuxiao Hu, Juncheng Liu, and Yuntian Chen. GenSR: Symbolic regression based on equation generative space. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8emIjwUQZg.

Chang Liu and Danial Chitnis. Eesizer: Llm-based ai agent for sizing of analog and mixed signal circuit. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 73(6):3825–3834, 2026. doi: 10.1109/TCSI.2025.3646359.

Chengjie Liu, Weiyu Chen, Anlan Peng, Yuan Du, Li Du, and Jun Yang. Ampagent: An llm-based multi-agent system for multi-stage amplifier schematic design from literature for process and performance porting, 2024. URL https://arxiv.org/abs/2409.14739.

Wenlong Lyu, Pan Xue, Fan Yang, Changhao Yan, Zhiliang Hong, Xuan Zeng, and Dian Zhou. An efficient bayesian optimization approach for automated optimization of analog circuits. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(6):1954–1967, 2018. doi: 10.1109/TCSI.2017.2768826.

Giang Ngo, Dat Phan Trong, Dang Nguyen, Sunil Gupta, and Svetha Venkatesh. Adaptive acquisition selection for bayesian optimization with large language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 149314–149336, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f1b7a29fe9759d4c9d23f78738e40b42-Paper-Conference.pdf.

Guanyuan Pan, Shuai Wang, Yugui Lin, Tiansheng Zhou, Pietro Liò, Zhenxin Zhao, and Yaqi Wang. Vlm-cad: Vlm-optimized collaborative agent design workflow for analog circuit sizing, 2026a. URL https://arxiv.org/abs/2601.07315.

Guanyuan Pan, Tiansheng Zhou, Jianxiang Zhao, Zhi Li, Yugui Lin, Bingtao Ma, Yaqi Wang, Pietro Liò, and Shuai Wang. Graph neural networks based analog circuit link prediction. *Engineering Applications of Artificial Intelligence*, 163:113035, 2026b. ISSN 0952-1976. doi: 10.1016/j.engappai.2025.113035. URL https://www.sciencedirect.com/science/article/pii/S0952197625030660.

J. Ramos and M. Steyaert. Three stage amplifier with positive feedback compensation scheme. In *Proceedings of the IEEE 2002 Custom Integrated Circuits Conference (Cat. No.02CH37285)*, pp. 333–336, 2002. doi: 10.1109/CICC.2002.1012833.

Keertana Settaluri, Ameer Haj-Ali, Qijing Huang, Kourosh Hakhamaneshi, and Borivoje Nikolic. Autockt: Deep reinforcement learning of analog circuit designs. In *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 490–495, 2020. doi: 10.23919/DATE48585.2020.9116200.

Aditya Shahane, Saripilli Swapna Manjiri, Ankesh Jain, and Sandeep Kumar. Graph of circuits with gnn for exploring the optimal design space. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 6014–6025. Curran Associates, Inc., 2023. doi: 10.52202/075280-0263. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/12da92b7c64176eb6eb6ad0ae31554fd-Paper-Conference.pdf.

Jinyi Shen, Zihao Chen, Ji Zhuang, Jiangli Huang, Fan Yang, Li Shang, Zhaori Bi, Changhao Yan, Dian Zhou, and Xuan Zeng. Atelier: An automated analog circuit design framework via multiple large language model-based agents. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 45(1):31–44, 2026. doi: 10.1109/TCAD.2025.3573228.

Zhuofu Tao, Yichen Shi, Yiru Huo, Rui Ye, Zonghang Li, Li Huang, Chen Wu, Na Bai, Zhiping Yu, Ting-Jung Lin, and Lei He. Amsnet: Netlist dataset for ams circuits. In *2024 IEEE LLM Aided Design Workshop (LAD)*, pp. 1–5, 2024. doi: 10.1109/LAD62341.2024.10691781.

Gemma Team et al. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Xinyue Wu, Fan Hu, Shaik Jani Babu, Yi Zhao, and Xinfei Guo. Easysize: Elastic analog circuit sizing via llm-guided heuristic search, 2025. URL https://arxiv.org/abs/2508.05113.

Zhe Ye, Zhengxu Yan, Jingxuan He, Timothe Kasriel, Kaiyu Yang, and Dawn Song. Verina: Benchmarking verifiable code generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 38933–38972, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/41b8c80f9113b9f8e2e129447221682a-Paper-Conference.pdf.

Shuhan Zhang, Fan Yang, Changhao Yan, Dian Zhou, and Xuan Zeng. An efficient batch-constrained bayesian optimization approach for analog circuit synthesis via multiobjective acquisition ensemble. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(1):1–14, 2022. doi: 10.1109/TCAD.2021.3054811.

Yang Zhao, Chengxiao Dai, Wei Zhuo, Yue Xiu, and Dusit Niyato. Clause: Agentic neuro-symbolic knowledge graph reasoning via dynamic learnable context engineering. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 82680–82706, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8598c5d05e1d045919962e430b0bdf0d-Paper-Conference.pdf.

A Extended related work

Analog circuit sizing has long been formulated as a black-box numerical search over device dimensions. Bayesian optimization and reinforcement learning reduce simulation demand in low-dimensional spaces, but they still rely on simulator evaluations as their primary feedback (Lyu et al., 2018; Ahmadzadeh et al., 2026; Zhang et al., 2022; Settaluri et al., 2020). Surrogate-assisted sizing replaces selected simulator calls with learned estimates (Lyu et al., 2018; Budak et al., 2022; Shahane et al., 2023), which require paired device-dimension and circuit-metric examples. Existing public circuit resources, however, focus mainly on topology and netlist structure (Li et al., 2025; Tao et al., 2024; Pan et al., 2026b; Bhandari et al., 2025; Gao et al., 2025), rather than providing the multi-size parameter-to-metric supervision needed to train sizing surrogates.

Language-model agents interpret netlists and steer analog sizing via natural-language reasoning. LMABO dynamically schedules acquisition functions within Bayesian optimization (Ngo et al., 2026), and VLM-CAD incorporates multimodal visual-textual reasoning into sizing workflows (Pan et al., 2026a). A parallel line derives explicit analytical equations instead of treating the simulator as a black box: White-Op constructs a behavioral symbolic reasoning loop that traces simulation-theory discrepancies back to analytical steps (Chen et al., 2026), NEMESIS synthesizes performance equations and refines them via SPICE-in-the-loop repair (Ghosh et al., 2026), and Bujana & Karsilayan (2026) extract operating-point parameters to iteratively correct model errors. While these works establish the value of symbolic guidance, they neither isolate the injection points as an experimental variable nor audit whether a prior from one environment remains valid on a distinct target testbench or a different circuit topology.

Beyond analog design, neuro-symbolic frameworks incorporate symbolic structures and domain constraints into neural models: EGG-SR injects equality constraints into symbolic regression (Jiang et al., 2026), GenSR navigates a structured equation space (Li et al., 2026), and VeriCoT enforces logical consistency across reasoning chains (Feng et al., 2026). Concurrently, an emerging line of work underscores the need to verify generated artifacts rather than trust them implicitly: VERINA benchmarks verifiable code generation against formal specifications (Ye et al., 2026), and CLAUSE grounds agentic reasoning in symbolic knowledge graphs (Zhao et al., 2026).

B Benchmark circuits

Table A1: Benchmark circuits, technology platforms, and target specifications. BSIM PTM stands for Berkeley Short-channel IGFET Model Predictive Technology Models (Cao, 2011), and SKY130 stands for the SkyWater SKY130 PDK (Edwards, 2020). — indicates that the metric is not measurable for the given circuit.

Circuit	Tech.	Specs	Gain (dB)	UGBW (MHz)	PM (°)	THD (dB)	Offset (mV)	Power (mW)
Miller Op-Amp	PTM 45 nm	SpecsA	> 54	> 1.00	> 60	< -60	< 5.0	< 0.15
		SpecsB	> 48	> 18.00	> 50	< -38	< 8.0	< 0.40
Discrete Op-Amp	Discrete MOSFET	SpecsA	> 90	> 0.35	> 60	< -55	< 0.1	—
		SpecsB	> 85	> 0.80	> 50	< -55	< 0.1	—
Class-AB Op-Amp	PTM 180 nm	SpecsA	> 65	> 1.00	> 65	< -26	< 0.5	< 0.60
		SpecsB	> 65	> 10.00	> 50	< -26	< 1.0	< 10.00
DFCFC	SKY130	SpecsA	> 90	> 3.00	> 50	< -60	< 0.5	< 2.00
		SpecsB	> 80	> 5.00	> 45	< -50	< 1.0	< 4.00
NMCNR	SKY130	SpecsA	> 90	> 2.00	> 55	< -50	< 1.0	< 1.00
		SpecsB	> 80	> 1.00	> 60	< -45	< 1.0	< 0.15
PFC	SKY130	SpecsA	> 80	> 5.00	> 50	< -50	< 1.0	< 1.50
		SpecsB	> 70	> 20.00	> 45	< -40	< 1.0	< 1.50

C Implementation details

Agent scaffold. All experimental configurations share a unified four-phase agent scaffold: 1) Schematic understanding, which parses the target netlist and specification constraints; 2) Initial

sizing, which synthesizes a complete starting candidate; 3) Iterative refinement, where the agent explores parameter updates under the symbolic operators described in Section 3.3; and 4) Handover, where a symbolic warm-start seed and the best verified refinement designs initialize the downstream TuRBO optimizer, which then refines the design under the same proposal cap.

Language model. All arms query Gemma-4-31B-it via the OpenRouter API. An optimization run executes a median of 34 to 42 language-model queries across arms and benchmark tasks.

Budget accounting. Each search chain runs at most 20 refinement iterations and at most 400 numerical proposals, and it stops early when a verified design passes the specification. Surrogate-screened proposals count toward this cap but do not use any simulator calls. The screening wrapper skips a proposal only when three conditions hold: its predicted reference cost exceeds $K\,J^*$ ($K = 2.0$), the cumulative skips stay below the number of verified simulations, and the proposal is not a probe. The wrapper simulates every 50th proposal regardless of its predicted cost. If a surrogate call fails, the wrapper disables itself and falls back to simulation.

One simulation call benchmarks one candidate across three Ngspice analyses: DC, AC, and transient. Three verification paths remain separate: normal simulation evaluates every committed design and gates every acceptance, the seed-transfer audit re-simulates the initial prior after the run as a post-hoc diagnostic (see Appendix H), and the pipeline has no automated pre-injection gate. Reported simulation totals include the upfront symbolic pre-computations, while the loop counter N_{eval} counts evaluations inside the search.

Cost and tolerance accounting. During the search the loop optimizes a strategy-weighted variant of the reference cost, in which an LLM-issued strategy sets a per-metric weight and exponent on a target-normalized shortfall: for metric k we compute $e_k = \max(0, s_k - m_k)/|s_k|$ for higher-is-better metrics and $e_k = \max(0, m_k - s_k)/|s_k|$ for lower-is-better metrics, then add $w_k(10e_k)^{p_k}$ with a per-metric weight w_k and exponent p_k (defaults $w_k = 1, p_k = 2$). Every reported comparison uses the static reference cost of Section 2, while the tolerance-adjusted pass rule decides acceptance.

We store THD as a positive percentage, convert it to dB as $20\log_{10}(\text{THD}/100)$, and treat non-positive values as unmeasured. We compare offsets in mV and power against $P_{\max}$ in W. The reported 5% tolerance applies in each metric's native scale: we relax a higher-is-better bound to $0.95\,s_k$, a lower-is-better positive bound to $1.05\,s_k$, and a negative dB bound (THD) to $0.95\,s_k$ toward acceptance. We re-scored the available per-run records (n = 70, 70, and 76) with a zero-tolerance predicate: strict Pass@1 becomes 5.7%, 20.0%, and 38.2% for no-symbolic, seed-only, and the full operator set, against 8.6%, 32.9%, and 55.3% at the 5% tolerance. These records include additional trials beyond the 60-run ablation sample, so the rates differ from Table 4.

Baselines. We evaluate the external baselines, EEsizer and VLM-CAD, using their original scaffolds and prompt templates under identical proposal caps and verification environments.

D STRUCTURAL CONSTANTS

Table A2: Constants of the sizing loop.

Constant	Value	Meaning
S	4	Stall threshold: consecutive iterations without cost improvement before injection
K	2.0	Screening multiple: proposals with predicted cost $> K\,J^*$ bypass simulation
τ_{skip}	$<$ verified evaluations	Skip budget: cumulative skips stay below the number of verified simulations
T_{probe}	50	Probe interval: we simulate every 50th proposal regardless of predicted cost
N_{agent}	20	Agent-refinement iterations before handover
N_{turbo}	400	Proposal cap for the numerical optimization stage
λ_k	$(1, 0.2, 0.1, 0.5, 10)$	Cost weights for (gain, UGBW, PM, THD, offset): we scale them so that any substantive violation pushes J above one

We summarize the hyperparameters and screening rules of the sizing loop (see Table A2). To construct a candidate design, the symbolic layer first computes the required input transconductance as $g_{m1} = 2\pi\,\text{UGBW} \cdot C_c$. It then scales transistor dimensions and bias currents rela-

tive to nominal baselines: input-stage widths scale by $r = g_{m1}/g_{m1}^{\text{base}}$ (we clip r to $[0.5, 8]$, with $g_{m1}^{\text{base}} = 3.4 \times 10^{-4}$ S), output-stage widths scale by $r_2 = g_{m2}/(10^{-3}\,\text{S})$, and bias-branch currents scale by $r_b = I_b/(30\,\mu\text{A})$, where we clip both r_2 and r_b to $[0.5, 4]$. Finally, the symbolic layer bounds all constructed transistor channel widths to $[2\,\mu\text{m}, 90\,\mu\text{m}]$.

E SURROGATE ARCHITECTURE AND TRAINING

Table A3: Surrogate accuracy under 5-fold cross-validation over the training rows, reported as the mean and standard deviation of R^2 across folds. Scale denotes the space in which we train and score each head.

Metric	Scale	$R^2 \uparrow$	Rows per fold
Gain	raw	0.737 ± 0.009	18,134
PM	raw	0.672 ± 0.005	18,134
DC output (auxiliary)	raw	0.735 ± 0.052	779
UGBW	log	0.414 ± 0.029	17,861
Power	log	0.308 ± 0.025	7,517
THD	log	0.847 ± 0.006	10,052
Offset	log	0.757 ± 0.191	779

Input and Target Preprocessing. The surrogate represents each candidate sizing design as a 140-dimensional vector spanning all unified parameter slots across the benchmark topologies, filling unassigned slots with empirical column means. We standardize these input features column-wise via $(X - \mu)/(\sigma + 10^{-9})$, where μ and σ denote the empirical mean and standard deviation of non-missing entries. For the regression targets, we apply a logarithmic transform to UGBW, power, THD, and offset before standardization, and clip all standardized targets z_k to $[-5, 5]$ to suppress destabilizing training outliers.

Neural Architecture. The surrogate comprises a total of 236,042 trainable parameters, organized as a shared backbone with seven prediction heads as follows:

- **Shared Trunk**: A 3-layer MLP mapping the 140-dimensional input to hidden representations. Each layer consists of Linear$(d_{\text{in}}, 256) \to$ SiLU $\to$ Dropout$(p = 0.1)$, with input dimension 140 in the first layer and 256 thereafter. The trunk omits LayerNorm and residual connections.
- **Individual Prediction Heads**: Seven single-layer linear projections Linear$(256, 1)$ independently predict each target metric and the auxiliary DC output voltage.
- **Auxiliary AC Refinement Head**: A 2-layer MLP structured as Linear$(256, 256) \to$ SiLU $\to$ Linear$(256, 3)$ that jointly predicts gain, UGBW, and PM.

We train the complete model end-to-end using the AdamW optimizer with a learning rate of 10^{-3} and weight decay of 10^{-4}, an element-wise masked MSE loss, a cosine schedule over 300 epochs, a batch size of 512, and early stopping with a patience of 12 epochs on an internal validation partition.

We report the predictive fidelity of the seven heads under 5-fold cross-validation over the training rows, with an internal split for early stopping (see Table A3).

E.1 SESSION-GROUPED SENSITIVITY

Table A4: Surrogate cross-validation sensitivity under session-grouped partitioning (mean $\pm$ standard deviation of R^2 across 5 folds). The row-level baseline corresponds to the default in-session evaluation reported in Table A3. **Bold** denotes the highest predictive fidelity per row.

Metric	Row-Level (Default)	Random Grouped	Stratified Grouped
Gain	$\mathbf{0.737 \pm 0.009}$	0.318 ± 0.341	0.285 ± 0.336
PM	$\mathbf{0.672 \pm 0.005}$	0.106 ± 0.330	0.193 ± 0.113
DC output	$\mathbf{0.735 \pm 0.052}$	0.421 ± 0.318	0.345 ± 0.375

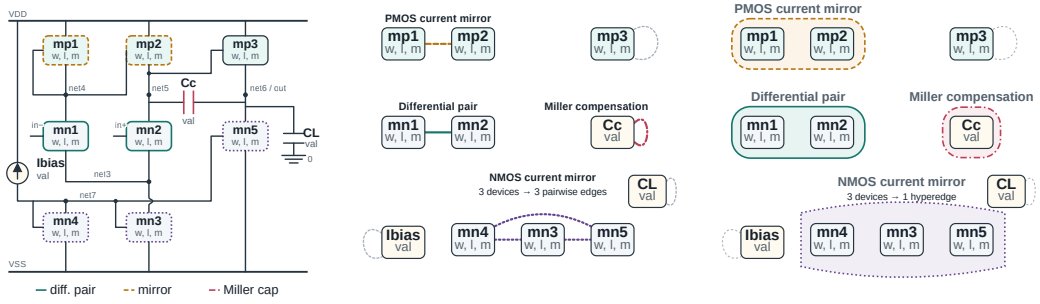
The default 5-fold cross-validation in Section 3.4 partitions data at the row level, allowing samples from the same optimization session to appear in both training and validation splits. To assess sensitivity to intra-session correlation and quantify potential data leakage, we repeat the evaluation using session-grouped 5-fold cross-validation across 183 independent search sessions under two partitioning strategies: 1) a uniformly random session split, and 2) a topology-stratified split that preserves balanced circuit representation across folds.

To mitigate destabilizing out-of-distribution extrapolation on unseen sessions, we clip predictions to the training target domain of $[-5, 5]$ standardized units. We exclude UGBW and power from this comparison, as their heavy-tailed raw-scale distributions induce extreme residual variance and unstable R^2 scores under session-level shifts. We report grouped results for the three metrics with stable scores and the row-level scores for all six metrics (see Table A3).

We compare the session-grouped results against the row-level reference (see Table A4). When we strictly hold out session boundaries, predictive accuracy declines substantially: stratified grouping yields a gain R^2 of 0.285 ± 0.336 and a PM R^2 of 0.193 ± 0.113. This degradation confirms that the surrogate functions primarily as an in-distribution ranking filter within local search trajectories, rather than an unconstrained cross-session predictor. This empirical limitation motivates the verification safeguards we introduce in Section 3.4: by bounding consecutive skips and enforcing periodic ground-truth SPICE probing, GLASS prevents downstream optimization from ever relying unchecked on out-of-distribution surrogate extrapolation.

F Structural Representation

We explored device-level structural representations of functional circuit blocks (see Fig. A1 for the pairwise and hypergraph encodings). A preliminary functional-block encoder built on this approach produced no accuracy gains over the multi-task surrogate and did not transfer across topologies, so GLASS does not use this encoder.



(a) Schematic of the two-stage Miller amplifier. The negative and positive inputs connect to net2 and net1, respectively.

(b) The graph projects the detected blocks into pairs. The three-device NMOS mirror forms three edges. Cc forms a self-loop.

(c) Functional hypergraph over the same devices. The three-device NMOS mirror forms one hyperedge instead of three edges.

Figure A1: Negative experiment: device-level graph representations of the two-stage Miller amplifier. Nodes represent circuit devices parameterized by their physical sizing attributes (W, L, M, or component value). Multi-device functional blocks decompose into pairwise clique edges in (b), but remain unified single hyperedges in (c) without edge distortion. Grey dashed loops indicate standalone devices not assigned to any multi-device sub-circuit rule.

G Extended Matched-Budget Compression

We report the detailed numerical results for the DFCFC SpecsA benchmark visualized in Fig. 2 (see Table A5). To establish the general operating boundary of symbolic warm-starting, we further extend this matched-budget protocol across five additional amplifier topologies under identical single-chain budgets, though at these chain counts the comparison is exploratory:

Table A5: Matched-budget search compression on DFCFC SpecsA under a fixed cap of 400 numerical proposals per chain. We retain chains that complete the full proposal cap or reach compliance early, so the arm sizes differ, and we drop no chain for failing to finish. Success requires full compliance, verified strictly via SPICE simulation. Median Evals denotes the evaluations required to first achieve compliance among successful chains. SPICE Evals reports mean simulator calls per chain. **Bold** denotes the best value per column.

Configuration	Chains	Successes	Rate (%) ↑	95% CI	Median Evals ↓	SPICE Evals ↓	Best Cost ↓
Vanilla search (a1)	19	5	26.3	[9.1, 51.2]	258	366	0.109
+ Symbolic warm-start (a2)	25	14	**56.0**	[34.9, 75.6]	148	292	0.103
+ Surrogate screening (a3)	12	2	16.7	[2.1, 48.4]	**62**	**180**	0.177
− Calibration (a4)	24	8	33.3	[15.6, 55.3]	216	350	**0.091**

- **Two-Stage Miller**: Under a 400-proposal cap, symbolic warm-starting increases the chain success rate from 44.4% (vanilla) to 50.0%.
- **Class-AB and Discrete Op-Amps**: Neither the vanilla arm nor the symbolic-assisted arms achieve the target specifications within the 400-proposal single-chain limit, reflecting the severe non-linearity of complementary input stages and discrete component search spaces.
- **NMCNR and PFC**: Under the same 400-proposal cap, no chain reaches the specification targets on SpecsA, confirming that single-chain scaling alone cannot overcome severe compensation pole shifts without multi-trial sampling ($Pass@k$).

H SEED-TRANSFER AUDIT DETAILS

Table A6: Seed-transfer audit for five SkyWater SKY130 task specifications. Predicted Gain denotes the analytical behavioral model output, while Measured Seed reports the initial SPICE simulation on the target verification testbench. The symbolic prior is deterministic: the five trials per topology share one seed. We compute Predicted Gain from the analytical model before the audit simulation.

Topology	Spec	Trials	Predicted Gain	Measured Seed	Transfer Status	Final Agent Gain
NMCNR	SpecsA	5	97.9 dB	97.9 dB	Valid (5/5)	96.9–114.9 dB
DFCFC	SpecsA	5	92.1 dB	92.1 dB	Valid (5/5)	94.2–106.3 dB
DFCFC	SpecsB	5	92.1 dB	92.1 dB	Valid (5/5)	89.7–109.7 dB
PFC	SpecsA	5	84.2 dB	−164.8 dB	Collapsed (0/5)	60.3–67.4 dB
PFC	SpecsB	5	84.2 dB	−164.8 dB	Collapsed (0/5)	43.6–73.2 dB

We document the seed-transfer audit across the audited SkyWater SKY130 amplifier tasks (see Table A6). Each entry records the behavioral model's predicted open-loop gain, the actual gain measured upon instantiating the seed directly into the target verification testbench, the transfer outcome, and the final gain range achieved by the optimization agent.

The audit substitutes the initial symbolic prior into each run's target AC testbench and re-simulates it with Ngspice. Control runs reproduce recorded outputs within 0.1%, validating the testbench. On NMCNR and DFCFC, the analytical model transfers with high precision, predicting 97.9 dB and 92.1 dB, matching the measured values across all 15 audited runs. On PFC, the prior predicts 84.2 dB but collapses to -164.8 dB in all 10 trials, as the uncalibrated bias-current mapping assigns $32.8\,\mu$A where usable operating points require about $5\,\mu$A. The agent partially recovers, increasing the open-loop gain to 43.6–73.2 dB. However, the final designs still fall short of the gain target on SpecsA, while the outcomes for SpecsB are mixed (see Table A6).

I PARAMETER-LEVEL EXAMPLE: MILLER OP-AMP

We provide a parameter-level trace of one benchmark run on the Miller Op-Amp SpecsA (see Table A7). Columns contrast the cold-start symbolic prior, the final verified design from the full GLASS framework, and the final design from an unguided (no-symbolic) run on the same specification and under the same verification protocol.

Table A7: Parameter-level case study on the Miller Op-Amp (45 nm PTM, SpecsA). Columns contrast the cold-start symbolic prior, the final verified design produced by GLASS, and the final design from the unguided (no-symbolic) baseline. We report widths (W) in μm and lengths (L) in nm, and M denotes device multiplicity. The verified gain satisfies specification within the 5% one-sided tolerance (see Section 4.2). All other metrics meet their targets directly.

	Symbolic prior	Verified (GLASS)	No-symbolic
Devices (W/L in μm/nm, multiplier M)			
mn1	27.6/120, 1	80/320, 1	450/937, 1
mn3	5.75/200, 2	5.75/951, 2	169/1599, 1
mn4	5.75/200, 1	5.75/398, 1	189/885, 1
mn5	6.88/200, 5	6.88/883, 5	301/941, 1
mp1	12.3/120, 1	12.3/320, 1	168/834, 1
mp3	72.3/120, 1	72.3/300, 1	490/775, 1
Passives and bias			
C_c (pF)	6.06	9	24
I_{bias} (μA)	43.1	25	30
C_L (pF)	10	10	10
Simulated performance (target)			
Gain (> 54 dB)	–	53.8	41.6
UGBW (> 1.0 MHz)	–	4.53	2.30
PM ($> 60^\circ$)	–	63.1	60.2
Power (< 0.15 mW)	–	0.146	0.145
THD (< -60 dB)	–	-75.1	-81.2
Offset (< 5 mV)	–	4.13	4.28

The analytical prior closely anticipates the final verified sizing: it matches the final verified design on five of the six transistor widths and preserves all device multiplicities (M). Consequently, closed-loop optimization primarily widens the differential input pair (`mn1`) and tunes channel lengths to achieve target DC gain. In contrast, the unguided search fails to exploit multi-finger scaling (retaining $M = 1$ across all devices), inflates total gate areas by an order of magnitude, and ultimately misses the target gain by 12.4 dB.

J WHY VERIFICATION BOUNDS PRIOR ERROR

Our claims for the cold-start operator π_{init} characterize the physical feasibility of the injected symbolic seeds rather than the trajectory of the subsequent agentic search. Because the sizing loop accepts candidates strictly upon SPICE verification (see Section 2), an inaccurate analytical prior can waste simulation budget but can never produce a false acceptance.

For each constrained metric $k \in \mathcal{K}$ we define an oriented performance margin $\mu_k(\mathbf{x})$ such that the specification target s_k is met exactly when $\mu_k(\mathbf{x}) \geq 0$. We write $\mu_k^{\text{ana}}(\mathbf{x})$ for the margin the calibrated behavioral model of Section 3.2 predicts and $\mu_k^{\text{sim}}(\mathbf{x})$ for the margin Ngspice measures.

Proposition 1 (Seed feasibility transfer)**.** *Assume that over the circuit operating domain the calibrated behavioral model predicts every constrained metric within an error bound ε_k:*

$$\left|\mu_k^{sim}(\mathbf{x}) - \mu_k^{ana}(\mathbf{x})\right| \leq \varepsilon_k, \qquad k \in \mathcal{K}, \tag{A1}$$

for all designs $\mathbf{x}$ reached by the loop. Let $\mathbf{x}_s$ be a symbolic seed whose analytic margins satisfy $\mu_k^{ana}(\mathbf{x}_s) \geq \varepsilon_k$ for every $k \in \mathcal{K}$. Then the simulated seed satisfies every target specification, i.e., $\mu_k^{sim}(\mathbf{x}_s) \geq 0$ for all $k \in \mathcal{K}$.

Proof. Fix an arbitrary metric $k \in \mathcal{K}$. Decomposing the simulated margin and applying the bound equation A1 yields:

$$\mu_k^{\text{sim}}(\mathbf{x}_s) = \mu_k^{\text{ana}}(\mathbf{x}_s) + \left(\mu_k^{\text{sim}}(\mathbf{x}_s) - \mu_k^{\text{ana}}(\mathbf{x}_s)\right) \geq \mu_k^{\text{ana}}(\mathbf{x}_s) - \varepsilon_k \geq 0, \tag{A2}$$

where the final inequality uses the margin condition. Because k is arbitrary, the seed satisfies all target specifications simultaneously. □

Corollary 1 (No false acceptance)**.** *Assume equation A1. If the margin condition fails for any metric k, the loop either verifies the candidate seed through simulation or rejects it at SPICE verification. In neither case does it accept an unverified or infeasible prior as a compliant solution.*

Proof. The loop accepts candidates only after re-simulation (see Section 2), and the reported design is the best verified design. A seed that violates the margin condition produces either a compliant measurement, which the loop admits as verified, or an out-of-specification measurement, which the loop discards. Consequently, every accepted design carries a verified physical margin. □

Remark 1 (Conditioning of the inversion map)**.** *We state bound equation A1 in metric space, which is the form we can check empirically. If instead we bound the model error in operating-parameter space, $\|\hat{\mathbf{p}} - \mathbf{p}\| \leq \eta$, then $\varepsilon_k \leq L_k \eta$, where L_k denotes the Lipschitz constant of metric k with respect to the operating parameters $\mathbf{p}$. Mapping operating parameters back to device dimensions relies on the two-stage inversion $\mathcal{T}$ of Section 3.2, whose Jacobian is $\mathbf{J}_{\mathcal{T}}$. The constant L_k then scales with the condition number $\kappa(\mathbf{J}_{\mathcal{T}}) = \|\mathbf{J}_{\mathcal{T}}\| \, \|\mathbf{J}_{\mathcal{T}}^{-1}\|$, so an ill-conditioned inversion amplifies analytical error and tightens the margin requirement.*

Remark 2 (Independence from bound tightness)**.** *We do not require ε_k to be small. The condition in Proposition 1 is sufficient rather than necessary, and GLASS relies on simulation verification rather than analytical conservatism: the loop intercepts a seed that fails the margin condition during re-simulation, which keeps π_{init} sound even when the behavioral model is coarse (see Section 4.5 for the empirical transfer rate across benchmarks).*

Remark 3 (Scope of the guarantee)**.** *Proposition 1 constrains the seed, not the search. It does not claim that the loop reaches a compliant design, that the surrogate predicts correctly, or that stall recovery and warm start preserve feasibility. Those operators act on candidates that the loop re-simulates before acceptance.*