TASK-AWARE QUBO ALLOCATION FOR MIXED-PRECISION QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixed-precision quantization requires discrete allocation of weight and activation bit-widths, followed by recovery of the selected network. We develop a task-aware quadratic unconstrained binary optimization (QUBO) surrogate with separate weight and activation profiles, a bit-operation (BOP) cost, and selected structural priors, then refine the resulting route through direct validation-based PROTES search. Across three public pretrained SIDD checkpoints with source-image-disjoint calibration and evaluation, the production simulated-annealing QUBO routes are within 0.01 dB of the strongest SFC routes in a retrospective 19-solver comparison. Refinement gives single-run static gains of 0.27–1.01 dB over QUBO and 0.39–1.34 dB over the HAWQ-style allocator at approximately 4.1% routed-layer BOPs, while remaining 0.39–0.87 dB below FP32. On a compact NAFBlock-based denoiser evaluated on a held-out patch split with shared source images, the refined route improves static PSNR over HAWQ-style by 0.21–2.46 dB at nearby costs across the three tightest requested targets. Recovery narrows these differences and can change route ordering, so we evaluate allocation quality separately before and after LSQ+. We report achieved cost, search variation and optimization expense; BOP reductions are analytical, and the reference inference path remains floating point.

1 INTRODUCTION

Mixed-precision quantization assigns different numerical precisions to network components to balance task quality and resource cost. Methods such as HAQ and HAWQ use hardware feedback or sensitivity estimates to guide allocation Wang et al. (2019); Dong et al. (2019; 2020). Learned quantizers, including LSQ and LSQ+ Esser et al. (2019); Bhalgat et al. (2020), adapt scales and offsets once precisions are fixed. The allocation and recovery problems are therefore related but distinct: a route that performs well before recovery need not remain the best route afterward.

A tractable allocation surrogate must also distinguish weights from activations. Local weight reconstruction error provides a candidate-specific signal, but intermediate activation perturbations affect subsequent computation. In the studied restoration network, unrestricted activation menus produce severely degraded routes under the tested QUBO search, motivating the empirical activation profile and asymmetric candidate space. Selected structural dependencies can further influence how these errors combine. These observations motivate separate sensitivity models and interaction terms rather than a layer-independent compression reward.

We combine a global discrete surrogate with a second stage that evaluates candidate networks directly. The first stage uses QUBO to trade normalized quantization damage against an analytical BOP cost. Hand-selected activation coupling and module-order damage terms incorporate structural priors. The second stage initializes PROTES, a tensor-train probabilistic optimizer, from the best observed QUBO route Batsheva et al. (2023); Oseledets (2011). A soft compute-growth penalty discourages refinement from purchasing quality through a large increase in precision.

Our contribution is a task-aware QUBO allocation model and an empirical assessment of when direct network search adds value. Separate activation profiling and structural priors make the surrogate useful for joint routing; QUBO-seeded PROTES tests the remaining gap to direct task evaluation. We separate allocation quality before recovery, quality after recovery, and measured execution behavior.

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.

The evaluation includes a HAWQ-style sensitivity–knapsack baseline, three public pretrained SIDD checkpoints, 19 QUBO solvers, and stochastic-search and validation-subset studies.

2 Related Work

Quantization and allocation. Post-training methods such as AdaRound and BRECQ optimize rounding or reconstruction without full retraining Nagel et al. (2020); Li et al. (2021). PACT, LSQ and LSQ+ learn quantizer parameters during training Choi et al. (2018); Esser et al. (2019); Bhalgat et al. (2020). HAQ uses reinforcement learning for allocation, while HAWQ variants exploit Hessian sensitivity Wang et al. (2019); Dong et al. (2019; 2020). Differentiable precision search, including EdMIPS, instead relaxes discrete choices Cai & Vasconcelos (2020); Wang et al. (2020). HAWQ-V3 further addresses integer-only execution and hardware-aware allocation Yao et al. (2021). Our evaluated HAWQ-style baseline combines Hessian weight sensitivity with the shared activation profile and a multiple-choice allocator. It evaluates this allocation recipe within our common quantization and recovery implementation. Uniform allocation isolates the value of layer-dependent precision, while the HAWQ-style allocator compares a Hessian-based allocation rule under the same activation profile and recovery stack. This shared stack controls implementation differences, but does not evaluate the complete HAWQ-V3 deployment system or establish superiority over differentiable search.

Discrete surrogates and direct search. QUBO represents binary allocation decisions and pairwise interactions Kochenberger et al. (2014); Glover et al. (2018); related neural-network compression formulations have also been studied Subiñas et al. (2025); Wang et al. (2026); Orabi et al. (2026). Muñiz Subiñas et al. optimize per-parameter rounding at a fixed precision, Wang et al. study joint pruning and quantization using adiabatic quantum computing, and Orabi et al. combine task-aware QUBO pruning with tensor-train refinement. Our binary decisions instead select paired weight and activation bit-widths while preserving the network structure; the surrogate additionally models BOP coupling and empirical activation damage, and the evaluation separates static allocation from post-recovery behavior. We primarily solve the surrogate with classical simulated annealing Kirkpatrick et al. (1983). We also benchmark tabu search Glover (1989), simulated quantum annealing Kadowaki & Nishimori (1998), coherent-Ising dynamics Tiunov et al. (2019); Reifenstein et al. (2021); Leleu et al. (2021), simulated-bifurcation variants Goto et al. (2019), and a classical software implementation of local quantum annealing (LQA) Bowles et al. (2022) through D-Wave samplers and the MindQuantum QAIA implementations Xu et al. (2024). PROTES optimizes a tensor-train sampling distribution using black-box function evaluations Batsheva et al. (2023); its use on QUBO-based machine-learning problems has also been studied by Salloum et al. Salloum et al. (2025). Our distinction is its use after a mixed-precision sensitivity-based QUBO allocation, with direct task evaluation and an explicit compute-growth penalty. The purpose of this combination is to correct surrogate mismatch; its usefulness must be assessed against both simpler allocation and unseeded search.

3 Sensitivity-Aware Allocation and Direct Refinement

Figure 1 summarizes the allocation and recovery stages. We first construct a sensitivity-based discrete surrogate, then refine its selected route using direct validation evaluation.

Candidate damage. For each routed convolution n, we select a weight precision $b \in \mathcal{C}^W = \{4, 5, 6, 7, 8\}$ and an input-activation precision $a \in \mathcal{C}^A = \{6, 7, 8\}$. Let $q^W_{n,b}$ and $q^A_{n,a}$ be binary one-hot variables. The candidate damage coefficients are

$$d^W_{n,b} = F^W_n \epsilon^W_{n,b}, \qquad d^A_{n,a} = F^A_n \epsilon^A_{n,a}. \tag{1}$$

Here ϵ^W and ϵ^A are tensor reconstruction MSEs under candidate quantization. F^W is a square-root/max-normalized exponential moving average of squared weight-gradient sums. It is Fisher-inspired, not an exact Fisher trace. F^A is a square-root/max-normalized score obtained by temporarily quantizing one convolution output and evaluating its final network effect. The final activation score differs from the local activation-gradient proxy used in preliminary Phase 2 (Appendix D.2).

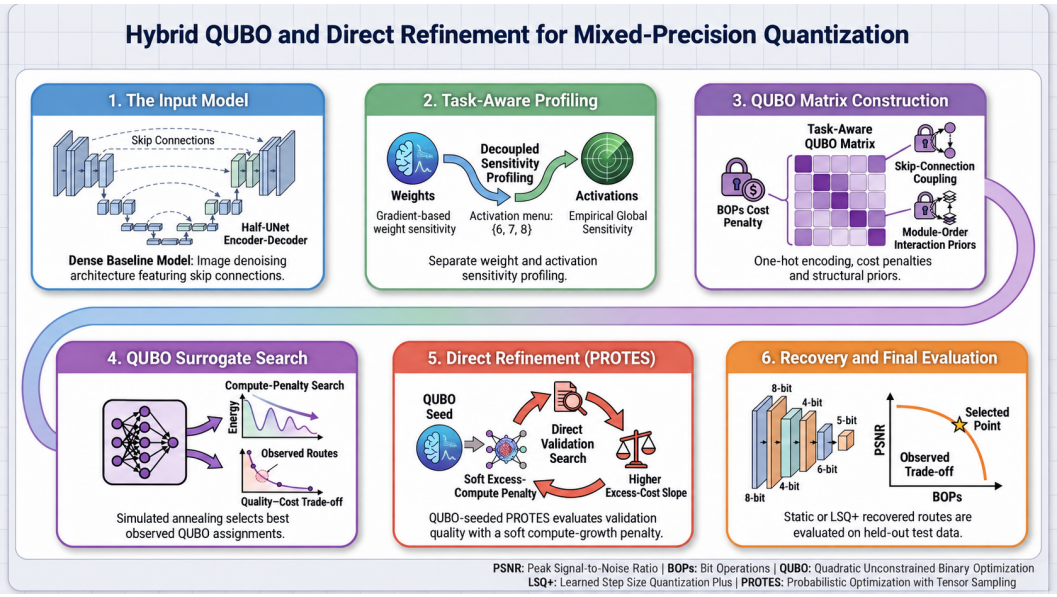


Figure 1: Hybrid QUBO allocation and direct validation refinement. The input network and trade-off sketches are schematic, not measured curves or a complete architectural specification. Weight and activation profiling, QUBO surrogate search, PROTES refinement and LSQ+ recovery are distinct stages; both search stages are heuristic. The activation menu shown is the final joint-quantization menu. Primary-model routes are evaluated on a held-out patch split whose source images are shared across partitions.

Candidate activation MSE instead samples the convolution input; these are associated but distinct tensor locations.

We preserve the profiling variants used by the experiments. Development all-convolution studies measure probe reconstruction MSE against ground truth. Additional-architecture studies measure disturbance from FP32 output using MSE for restoration and KL divergence for classification. Development weight candidate errors use a scale grid, while the reconstructed QUBO and additional architectures use isolated 50-step scale calibration. Neither is network-level LSQ+ recovery. Appendix A specifies the normalization, quantizers and variants in detail.

Choice of activation profile. Preliminary studies used shared weight-gradient scores, then separate local activation gradients (Phase 2). Neither the latter nor an activation scalar resolved degraded routes. The final design uses downstream empirical profiling and an asymmetric menu. Tables 5 and 6 summarize these comparisons; Appendix D.2 retains their chronology and protocols.

QUBO objective. Collecting the binary variables into q, the surrogate combines

$$E(q) = D(q) + \gamma C(q) + \omega R(q) + \alpha P(q), \tag{2}$$

where D sums candidate damage after common maximum normalization, C is the weight–activation compute coupling, R combines normalized activation-damage products for consecutive convolutions in module traversal order, and P contains one-hot and selected activation-mismatch penalties. The default damage scale is one and $\omega = 0.15$ in the main formulation. For the reported main configuration, $\omega = 0.15$, the activation menu and the exclusion rules are treated as fixed preliminary design choices. The later test-loader sweeps in Appendix D.2 are retrospective diagnostics and are not used to select the main results. Equation 2 groups the implemented matrix contributions; Appendix A gives their exact coefficient construction. The one-hot penalty is equivalent, up to a constant, to $\sum_g (\sum_{u \in g} q_u - 1)^2$. Selected mismatched activation choices receive an additional exclusion penalty.

Matrix assembly. For an upper-triangular QUBO, initialize $Q_{uu} = \widehat{d}_u - \alpha$, where $\widehat{d}_u = d_u / \max_v d_v$. Add 2α to competing one-hot pairs and selected skip-mismatch pairs, γC_{uv} to weight–activation cost pairs, and $\omega \widehat{d}^A_u \widehat{d}^A_v$ to module-order activation pairs. Contributions to the same pair sum; all other entries are zero. Appendix A specifies the adaptive exclusion scale and coefficient definitions.

These are hand-selected priors: coupled convolution-input variables proxy fusion paths, and module traversal need not follow data-flow edges. They provide neither graph-derived interactions nor an integer fusion interface.

We anneal the surrogate and retain the best observed route. The extended implementation repairs one-hot selections and performs up to four coordinate-improvement passes; preliminary raw-solver studies remain a separate variant. Gamma search seeks a requested compute level but discrete routes can miss it. The normalized routed-layer cost is

$$B(\mathcal{A}) = \frac{\sum_{n \in \mathcal{R}} \mathrm{MAC}_n b^W_n b^A_n}{32^2 \sum_{n \in \mathcal{R}} \mathrm{MAC}_n}, \tag{3}$$

with routed convolutions $\mathcal{R}$. Protected convolutions and other floating-point operations are excluded. BOP percentages below refer to this proxy, not whole-network latency or energy.

QUBO-seeded PROTES. PROTES samples candidate routes from a tensor-train probability model initialized around the QUBO route. Restoration candidates are evaluated on 32 validation patches. With seed cost B_0, refinement maximizes

$$S(\mathcal{A}) = M(\mathcal{A}) - \gamma B(\mathcal{A}) - (h - \gamma) \max\{B(\mathcal{A}) - B_0, 0\}, \quad h = \max\{10\gamma, 100\}, \tag{4}$$

where M is validation PSNR or classification accuracy in percentage points. The continuous score charges excess compute at the higher rate; its soft penalty permits routes above the seed cost. The final route is evaluated on the held-out patch split with shared source images, separately before and after LSQ+. Full-sweep and follow-up experiments use different budgets and cache representations; the appendix reports them separately.

The excess-cost term raises the marginal compute charge from γ to h above the seed cost, while retaining the original charge for cheaper routes. In preliminary runs, a small linear penalty allowed substantial compute expansion; the added tax encouraged quality gains at similar or lower cost (Table 7). This motivates the continuous score rather than a hard feasibility constraint.

4 Experimental Protocol

Evidence scope. The primary sweep reports individual recoveries; repeated-search studies measure search variation conditional on a fixed recovery seed. Static all-convolution ablations test formulation and solver sensitivity. Where recovery is reported for additional models, it uses one recovery replicate; the public-checkpoint benchmark is static only. Table 4 maps these variants to their splits and routing settings; results are interpreted within each protocol.

The primary model is a custom width-32 HalfUNet denoiser using NAFBlocks Chen et al. (2022) and additive multi-scale fusion. SIDD Small provides 160 RGB noisy/clean image pairs Abdelhamed et al. (2018). Four quadrants per image are resized to 256×256, producing 640 patches, then split into 512/64/64 train/validation/test patches with seed 42. Training exposes three accesses per patch with flip augmentation. The primary batch size is 16. Because splitting follows patch extraction, source images can cross partitions. All reported primary SIDD scores use this resized-patch evaluation.

The main comparison retains full precision at the input/output boundaries by protecting the first and last convolutions. It routes 38 internal convolutions and uses 304 binary variables. The all-convolution statistical and objective-development studies also route those boundary layers, giving 40 convolutions and 320 variables. Table 4 identifies the setting used by each study.

The pretrained benchmark instead uses public NAFNet-width32 and Baseline-width32 checkpoints Chen et al. (2022) and a public HINet-0.5x checkpoint Chen et al. (2021) on the official SIDD validation blocks. Five of 40 source images (160 patches) are reserved for calibration and

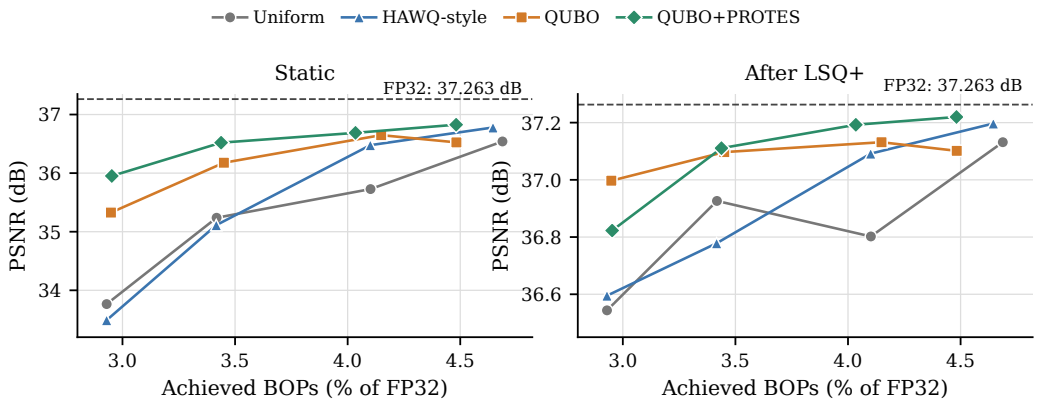


Figure 2: HalfUNet comparison across the four requested BOP targets for which all methods are reported in Table 1. Markers show recorded evaluations at achieved routed-layer BOPs; thin lines connect the requested-target sequence only as a visual guide and are not a certified Pareto frontier. The left panel reports static evaluation and the right panel the corresponding LSQ+ recovery.

search; the other 35 (1,120 patches) are used only for final evaluation. The first and last convolutions remain FP32, as do HINet transposed convolutions. These runs use static fake quantization without LSQ+ recovery. The solver study reuses the resulting fixed QUBOs and adds the protected 38-layer HalfUNet problem. Except for this retrospective solver benchmark, every QUBO assignment reported elsewhere in the paper—including the primary operating points, figures, robustness studies and PROTES initializations—was produced by simulated annealing under the budget stated for that experiment. Thus, Table 3 evaluates solver choice retrospectively; it does not silently replace any earlier assignment.

Static weights use a minimum-MSE symmetric scale search: the preliminary evaluator uses an 80-point grid and the additional experiments use 64 points. Table 4 preserves these evaluation settings. Activations use dynamic symmetric quantization. Convolutions, biases, accumulation, normalization and residual operations remain floating point. Extended LSQ+ recovery uses four quantizer-only epochs followed by fourteen epochs updating quantizers and convolution parameters, capped at 100 batches per epoch. Restoration uses L1 loss and classification cross-entropy.

Restoration outputs are clipped to $[0, 1]$. The shared evaluator averages PSNR computed from each batch's MSE and reports SSIM Wang et al. (2004). The paired statistical analysis instead computes per-patch PSNR. Dynamic activation scales include the batch dimension, so batch size and aggregation are part of the evaluation protocol. Routing and calibration use validation data.

5 Results

We report quality before and after recovery, the incremental benefit and cost of direct search, structural ablations, and measured execution behavior. The primary compute comparisons use achieved routed-layer BOPs; preliminary approximate costs and missing costs are identified explicitly.

5.1 Allocation Quality Before and After Recovery

Table 1 condenses the six-target sweep. At the two loosest targets it reports only the distinct uniform baselines because each non-uniform method reused exactly the same route already shown in the 4.688% block. Figure 2 compares achieved quality and compute for the four fully populated target groups. Static and recovered records are paired by assignment ID.

At the requested 4.102% target, refinement achieves 37.192 dB at 4.035% BOPs, versus 37.132 dB at 4.149% for QUBO and 37.092 dB at 4.101% for the HAWQ-style allocator. At the requested 3.418% target, QUBO reaches 37.097 dB at 3.450%, compared with 36.778 dB at 3.417% for HAWQ-style. These are nearby-cost observations, not exact matched-cost significance tests.

Table 1: Primary patch-random HalfUNet comparison. Costs are routed-layer BOP percentages relative to FP32; quality cells give PSNR (dB) / SSIM. The two loosest requested targets show only their distinct uniform routes; all methods are shown from 4.688% downward. The omitted non-uniform rows at 6.250% and 5.469% repeat the same assignments shown at 4.688% (their separate recovery evaluations remain in Figure 2). Bold marks each metric maximum within a displayed target block, excluding FP32.

Method	Achieved BOPs (%)	Static PSNR / SSIM	LSQ+ PSNR / SSIM
FP32 reference	100.000	37.263 / 0.9328	–
Uniform	6.250	36.877 / 0.9245	37.242 / 0.9316
Uniform	5.469	36.919 / 0.9254	37.196 / 0.9308
Uniform	4.688	36.540 / 0.9199	37.132 / 0.9305
HAWQ-style	4.645	36.779 / 0.9224	37.197 / 0.9299
QUBO	4.483	36.525 / 0.9178	37.101 / 0.9295
QUBO + PROTES	4.482	**36.826 / 0.9237**	**37.220 / 0.9309**
Uniform	4.102	35.727 / 0.9026	36.802 / 0.9242
HAWQ-style	4.101	36.477 / 0.9152	37.092 / 0.9278
QUBO	4.149	36.648 / 0.9215	37.132 / 0.9298
QUBO + PROTES	4.035	**36.687 / 0.9221**	**37.192 / 0.9308**
Uniform	3.418	35.238 / 0.8936	36.926 / 0.9268
HAWQ-style	3.417	35.112 / 0.8810	36.778 / 0.9215
QUBO	3.450	36.177 / 0.9102	37.097 / 0.9292
QUBO + PROTES	3.438	**36.520 / 0.9193**	**37.111 / 0.9294**
Uniform	2.930	33.765 / 0.8552	36.544 / 0.9178
HAWQ-style	2.930	33.490 / 0.8464	36.595 / 0.9175
QUBO	2.949	35.327 / 0.9058	**36.997 / 0.9279**
QUBO + PROTES	2.953	**35.952 / 0.9113**	36.823 / 0.9239

Recovery changes the ordering of routes. At the tightest target, refinement improves static PSNR from 35.327 to 35.952 dB but reduces recovered PSNR from 36.997 to 36.823 dB. The requested 6.250%, 5.469% and 4.688% targets reuse the same QUBO route and the same refined route, with separate recovery evaluations. The discrete gamma-to-route map did not produce distinct non-uniform assignments near the two loosest targets; this limits control of the allocator at high budgets. Uniform quantization is competitive at larger costs. The original uniform recovery score at 4.102% is 36.802 dB; a later run on the same assignment records 37.010 dB. The difference illustrates why isolated post-recovery gaps should be scoped as observations rather than reliable method-level effects.

5.2 Pretrained SIDD Denoisers

Figure 3 evaluates the same allocation methods on three public pretrained checkpoints. These are training-free static results; calibration/search and evaluation use disjoint source images. Full BOP and SSIM records appear in Appendix Table 10.

Refinement improves over QUBO by 1.012 dB on NAFNet, 0.272 dB on the NAFNet baseline, and 0.538 dB on HINet, while remaining 0.389, 0.398 and 0.870 dB below the corresponding FP32 checkpoints. The methods have close analytical costs, but not identical assignments. The result supports a consistent static refinement benefit across these architectures; it does not predict post-LSQ+ ordering. Post-recovery generalization beyond the primary HalfUNet remains an open question because these public checkpoints were evaluated without LSQ+.

5.3 Does QUBO Initialization Help Direct Search?

We compare random search, randomly initialized PROTES, network-score simulated annealing and QUBO-seeded PROTES on one 32-patch validation subset, with seeds 1701–1703 and a nominal budget of 2,000 proposals. We also recover and evaluate the fixed QUBO seed; its recovery eval-

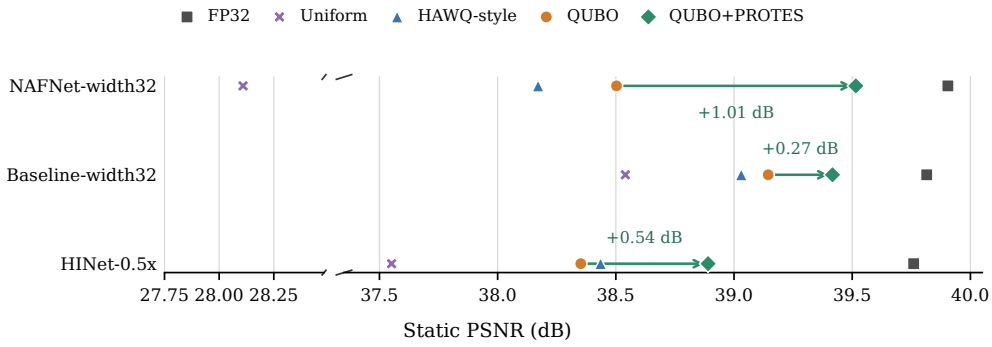


Figure 3: Training-free static PSNR on three public pretrained SIDD checkpoints. Arrows show QUBO-to-PROTES changes. Non-FP32 methods target 4.102% routed-layer BOPs; achieved costs are 3.999–4.106%. The broken horizontal axis retains the 28.109 dB NAFNet uniform result while keeping the remaining routes legible. Exact PSNR, SSIM and cost values appear in Appendix Table 10.

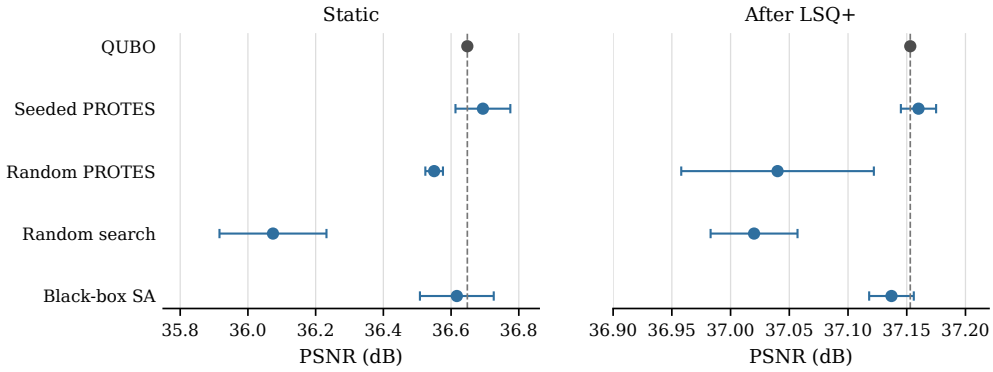


Figure 4: Search comparison at one requested compute target on the held-out patch split with shared source images. Points and error bars are means and standard deviations over three search seeds; fixed QUBO is a single assignment and is marked by the dashed line in each panel. LSQ+ variation is conditional on one fixed recovery seed. The tighter horizontal scale after LSQ+ makes the reduced separation among methods visible.

uation is separate from the six-target sweep. Figure 4 visualizes quality before and after recovery; Appendix Table 15 gives the exact scores, achieved mean costs and realized evaluation counts.

Seeded PROTES has the highest mean static score among the tested stochastic methods, but its improvement over QUBO alone is 0.046 dB before LSQ+ and 0.007 dB afterward. Achieved costs also differ slightly. The mean realized unique evaluations are approximately 1,953 for seeded PROTES, 2,000 for random PROTES and random search, and 1,700 for network-score SA. These are comparable nominal budgets, not identical evaluation counts.

5.4 Effect of Structural Priors

Table 2 reports preliminary structural comparisons at $\gamma = 50$ and 200. At $\gamma = 50$, coupling raises PSNR from 35.105 to 35.386 dB while increasing cost from 3.603% to 4.157%. At $\gamma = 200$, adding module-order interactions improves static PSNR by 0.446 dB relative to activation coupling alone at the same reported cost. The uncoupled comparison also changes cost. These single-run preliminary observations are separate from the protected-layer recovered comparison; Appendix D.3 retains the wider sweep.

Table 2: Development structural-prior comparisons before LSQ+. Average bits are preliminary layer averages. Only the two coupled $\gamma = 200$ rows have identical reported BOPs. Single-run records do not establish recovered multi-seed effects or an optimal interaction coefficient.

γ	Formulation	Avg. W / A	BOPs (%)	PSNR	SSIM
50	No structural priors	5.58 / 7.17	3.603	35.105	0.887
50	Activation coupling	6.05 / 7.00	4.157	35.386	0.894
200	No structural priors	5.90 / 7.12	3.568	34.368	0.869
200	Activation coupling	5.53 / 7.05	3.374	34.455	0.870
200	Coupling + $\omega = 0.15$	5.55 / 7.00	3.374	34.901	0.891

Table 3: Production SA/QUBO and representative backends on the three public pretrained checkpoints. Entries are PSNR (dB) / achieved BOPs (%). Energy rank is computed from best-energy trials across all four fixed QUBOs; "Budget" counts the displayed closest-to-target assignments within 0.25 percentage points of 4.102%. The standardized SA row is the controlled 16-read benchmark, not the production pipeline. NMFA's larger scores coincide with budget overshoot on NAFNet and Baseline.

Solver	Energy rank	Budget	NAFNet	Baseline	HINet
Production SA/QUBO	–	3/3	38.503 / 4.101	39.145 / 4.104	38.352 / 4.098
SFC	1.25	3/3	38.502 / 4.091	39.135 / 4.057	38.352 / 4.091
LQA	1.50	3/3	38.502 / 4.095	39.140 / 4.079	38.353 / 4.089
NMFA	3.00	1/3	38.618 / 4.660	39.277 / 4.646	38.480 / 4.204
PulseCIM	18.75	3/3	38.361 / 4.014	38.826 / 4.046	38.020 / 4.001
Standardized SA	15.00	3/3	35.515 / 4.101	37.456 / 4.114	36.373 / 4.108

Formulation development. Appendix D.2 preserves the preliminary weight-only, joint-allocation and refinement-objective comparisons that motivated the final activation profile and excess-cost penalty. Because their protocols evolved, we do not present them as equal-setting main comparisons.

5.5 Robustness of Profiling and Routing

Supporting studies identify several limits of stability. Five overlapping 32-patch validation subsets produce static scores of approximately 36.712 ± 0.037 dB, but only 37–58% pairwise agreement in both layer precisions. Probe-bit ranking correlations are high (Spearman 0.936–0.984), but this does not establish stable coefficient magnitudes or routes. The corrected all-convolution solver-seed study has PSNR standard deviations above 1 dB at two tested gamma values despite relatively stable surrogate energy.

Across four fixed QUBOs and five trials per backend, the 19-solver benchmark associates the classical SFC and LQA software backends with mean best-trial energy ranks of 1.25 and 1.50. More importantly, their selected routes match the production simulated-annealing QUBO routes on the three public checkpoints: SFC differs by only -0.001, -0.009 and -0.000 dB on NAFNet, Baseline and HINet, respectively. The much weaker standardized SA row uses 16 reads and 2,000 sweeps without the production gamma search, 500-read final solve and coordinate improvement; it is a controlled backend diagnostic, not the SA pipeline used for the paper's main QUBO results. Median solve-only times are 1.23 s for SFC, 1.50 s for LQA and 0.18 s for standardized SA. The benchmark also includes PulseCIM (Mifsud Bonici & Salloum, 2026). Appendix D.7 gives all backends, timings and network scores.

Scope of the solver result. The backend benchmark does not show that the main results are understated. Instead, it shows that the production SA pipeline already reaches essentially the same selected routes as the lowest-energy SFC/LQA backends on the three public checkpoints. The standardized SA control demonstrates sensitivity to solver budget and postprocessing, but should not be

used as a proxy for the production search. Lower QUBO energy also remains an imperfect predictor of network quality, as the model-specific failures in Appendix Figure 6 illustrate.

The all-convolution PROTES extension provides paired evidence at $(\omega, \beta, \gamma) = (0.15, 1, 85.77)$. At 3.140% BOPs, PROTES obtains 36.375 dB mean per-patch PSNR, compared with 34.994 dB for its 3.696% QUBO seed and 32.962 dB for Uniform W4/A8 at 3.125%. The paired gains are 1.382 dB (95% CI [1.235, 1.528]) and 3.413 dB (95% CI [3.095, 3.730]); both paired t-tests have $p < 2 \times 10^{-27}$. Shared source images and batch-dependent scales make these nominal patch-level, static results rather than scene-level or post-recovery significance claims.

Alternative Hessian-QUBO, entropy-based activation unlocking and sub-layer routing designs did not consistently improve the evaluated configurations (Appendix D.20). These observations motivate the simpler final design without establishing that those alternatives are universally inferior.

5.6 Optimization Cost and Reference Implementation

At the primary 4.102% target, the recorded end-to-end cost rises from 198.1 s for QUBO to 734.5 s for QUBO+PROTES; the separate 2,000-proposal study averages 1,249.8 s for refinement alone. The fake-quantized reference path executes floating-point convolutions: on an H100, batch-one inference is 3.002 ± 0.094 ms for FP32 and 6.735 ± 0.249 ms for QUBO, while serialized size is unchanged. Native low-bit kernels and packed tensors are required to realize the analytical savings. Appendix D.18 and Appendix D.10 give the full timing and deployment records.

6 Discussion and Limitations

Recovery changes the return on allocation search. QUBO provides a strong starting route in the primary repeated-search experiment, where LSQ+ leaves little additional gain for direct refinement. By contrast, all three public pretrained checkpoints show static gains from refinement, while post-recovery generalization on those models remains unmeasured. QUBO alone is therefore a practical choice when recovery is available and search budget is limited; direct refinement is most useful when static or recovered quality gaps remain. A direct study of surrogate scores and recovered route rankings would further characterize recovery-aware allocation.

The primary split shares source images, most recovery comparisons use one run, and batch-dependent activation scales make batch size part of the protocol. The shared stack includes only a HAWQ-style baseline; differentiable methods and exact MCKP/MIQP references remain untested. All-convolution and test-sweep diagnostics use separate protocols. The solver study reports heuristic best-observed solutions rather than certified optima: SFC/LQA match production SA on the public checkpoints, while weak standardized-SA settings degrade it.

The structural interactions are selected proxies. Routed-layer BOPs exclude protected layers and nonconvolutional operations, and ideal packed storage is not realized by the reference implementation. Native low-bit kernels and integer fusion are required for deployment gains; HAWQ-V3 demonstrates this dependence on supported kernels and hardware Yao et al. (2021).

7 Conclusion

Task-aware QUBO provides joint weight–activation routes, and PROTES improves static quality on three public denoisers. On HalfUNet, LSQ+ narrows the search gap and can reorder routes; recovery on the other models is unmeasured. Production SA matches SFC/LQA route quality on the public checkpoints. Solver budget, recovery and optimization expense therefore belong beside quality and achieved BOPs in the allocation decision.

REPRODUCIBILITY STATEMENT

Appendix A gives the implemented QUBO coefficient construction, and Appendix D records the routing variants, data partitions, candidate spaces, search budgets, seeds, achieved costs and static or recovered status of each result family. The protocol table explicitly separates the 38-convolution primary experiments from the 40-convolution statistical studies and the public-checkpoint benchmark. Numerical claims were checked against the saved CSV, JSON and notebook outputs from the corresponding experiments. The solver table reports all tested backends and distinguishes solve-only timing from network-evaluation cost.

AI USE STATEMENT

Generative AI tools assisted with software-code editing, organizing and interpreting saved experimental outputs, literature discovery, discussion of experiment design and methodology presentation, result interpretation, reference formatting, and manuscript drafting and editing. The authors independently checked numerical claims against saved CSV, JSON and notebook outputs, reviewed the cited literature and compiled manuscript, and take responsibility for the final content, including AI-assisted text, code and analysis.

REFERENCES

Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1692–1700, 2018.

Anastasiia Batsheva, Andrei Chertkov, Gleb Ryzhakov, and Ivan Oseledets. Protes: probabilistic optimization with tensor sampling. *Advances in Neural Information Processing Systems*, 36: 808–823, 2023.

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 696–697, 2020.

Joseph Bowles, Alexandre Dauphin, Patrick Huembeli, José Martinez, and Antonio Acín. Quadratic unconstrained binary optimization via quantum-inspired annealing. *Physical Review Applied*, 18 (3):034016, 2022.

Zhaowei Cai and Nuno Vasconcelos. Rethinking differentiable search for mixed-precision neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2349–2358, 2020.

Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. HINet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 182–192, 2021.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 293–302, 2019.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Fred Glover. Tabu search—part i. *ORSA Journal on Computing*, 1(3):190–206, 1989.

Fred Glover, Gary Kochenberger, and Yu Du. A tutorial on formulating and using qubo models. *arXiv preprint arXiv:1811.11538*, 2018.

Hayato Goto, Kosuke Tatsumura, and Alexander R. Dixon. Combinatorial optimization by simulating adiabatic bifurcations in nonlinear hamiltonian systems. *Science Advances*, 5(4):eaav2372, 2019.

Tadashi Kadowaki and Hidetoshi Nishimori. Quantum annealing in the transverse ising model. *Physical Review E*, 58(5):5355–5363, 1998.

Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Gary Kochenberger, Jin-Kao Hao, Fred Glover, Mark Lewis, Zhipeng Lü, Haibo Wang, and Yang Wang. The unconstrained binary quadratic programming problem: a survey. *Journal of combinatorial optimization*, 28(1):58–81, 2014.

Timothée Leleu, Farad Khoyratee, Timothée Levi, Ryan Hamerly, Takashi Kohno, and Kazuyuki Aihara. Scaling advantage of nonrelaxational dynamics for high-performance combinatorial optimization. *Communications Physics*, 4(1):266, 2021.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

Maximilian Mifsud Bonici and Hadi Salloum. PulseCIM: A preconditioned lookahead simulated coherent ising machine. *Available at SSRN 6446548*, 2026. doi: 10.2139/ssrn.6446548.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, pp. 7197–7206. PMLR, 2020.

Osama Orabi, Artur Zagitov, Hadi Salloum, Viktor A. Lobachev, Kasymkhan Khubiev, and Yaroslav Kholodov. Neural network pruning via QUBO optimization. *arXiv preprint arXiv:2604.05856*, 2026. doi: 10.48550/arXiv.2604.05856.

Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

Sam Reifenstein, Satoshi Kako, Farad Khoyratee, Timothée Leleu, and Yoshihisa Yamamoto. Coherent ising machines with optical error correction circuits. *Advanced Quantum Technologies*, 4 (11):2100077, 2021.

Hadi Salloum, Osama Orabi, Salman Ahmdi, Yaroslav Kholodov, Xiao Yu, Witold Pedrycz, and Mirko Farina. Black-box optimization of QUBO-based machine learning models using tensor train sampling. *TechRxiv preprint*, 2025. doi: 10.36227/techrxiv.175296816.62467576/v1.

Sergio Muñiz Subiñas, Manuel L González, Jorge Ruiz Gómez, Alejandro Mata Ali, Jorge Martínez Martín, Miguel Franco Hernando, and Ángel Miguel García-Vico. Optimization of the quantization of dense neural networks from an exact qubo formulation. *arXiv preprint arXiv:2510.16075*, 2025.

Egor S. Tiunov, Alexander E. Ulanov, and A. I. Lvovsky. Annealing by simulating the coherent ising machine. *Optics Express*, 27(7):10288–10295, 2019.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8612–8620, 2019.

Ying Wang, Yadong Lu, and Tijmen Blankevoort. Differentiable joint pruning and quantization for hardware efficiency. In *European Conference on Computer Vision*, pp. 259–277. Springer, 2020.

Zhehui Wang, Benjamin Chen Ming Choong, Tian Huang, Daniel Gerlinghoff, Rick Siow Mong Goh, Cheng Liu, and Tao Luo. Is quantum optimization ready? an effort towards neural network compression using adiabatic quantum computing. *Future Generation Computer Systems*, 174: 107908, 2026.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Xusheng Xu, Jiangyu Cui, Zidong Cui, Runhong He, Qingyu Li, Xiaowei Li, Yanling Lin, Jiale Liu, Wuxin Liu, Jiale Lu, et al. MindSpore Quantum: A user-friendly, high-performance, and ai-compatible quantum computing framework. *arXiv preprint arXiv:2406.17248*, 2024.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael Mahoney, and Kurt Keutzer. HAWQ-V3: Dyadic neural network quantization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 11875–11886. PMLR, 2021.

A Detailed QUBO Formulation

The proposed framework follows a sequential progression. First, layer sensitivity and candidate quantization errors are measured. These quantities are incorporated into a QUBO objective containing one-hot routing penalties, weight–activation computational-cost coupling, skip-connection compatibility terms, and adjacent-activation interaction terms. The best observed QUBO configuration is subsequently used to initialize PROTES.

A.1 Decision Variables and One-Hot Encoding

Let $\mathcal{C}^W$ denote the candidate weight bit-widths and $\mathcal{C}^A$ denote the candidate activation bit-widths. For each selected layer n, binary variables are defined as

$$q^W_{n,b} \in \{0,1\}, \qquad \forall b \in \mathcal{C}^W, \tag{5}$$

$$q^A_{n,a} \in \{0,1\}, \qquad \forall a \in \mathcal{C}^A. \tag{6}$$

The selected precisions are represented by

$$b^W_n = \sum_{b \in \mathcal{C}^W} b q^W_{n,b}, \qquad b^A_n = \sum_{a \in \mathcal{C}^A} a q^A_{n,a}. \tag{7}$$

The implemented QUBO assigns a large exclusion penalty to competing precision states within the same routing group. For every conflicting one-hot pair (i, j),

$$Q_{ij} \mathrel{+}= 2\alpha. \tag{8}$$

The diagonal coefficient for each precision state contains a corresponding negative exclusion contribution.

A.2 Task-Aware Weight Quantization Damage

For weight tensor W_n and candidate weight precision b, the symmetric quantization error is obtained using a grid search over the quantizer scale:

$$\epsilon^W_{n,b} = \min_{s \in \mathcal{S}_{n,b}} \frac{1}{|W_n|} \left\| Q_b(W_n; s) - W_n \right\|_2^2 . \tag{9}$$

For $b > 1$, the reference scale is

$$s_0 = \frac{\max |W_n|}{2^{b-1} - 1}, \tag{10}$$

and candidate scales are generated by multiplying s_0 by factors uniformly spaced between 0.05 and 1.20.

The preliminary grid-MSE formulation estimates candidate errors using a 64-point scale grid. The reconstructed QUBO, HAWQ-style candidate construction and new architecture experiments instead use an LSQ-calibrated weight-error variant: a temporary scale is optimized for 50 Adam steps at learning rate 10^{-2}, keeping the weight tensor fixed, and its final MSE replaces the grid-search error. This local calibration is distinct from network-level LSQ+ recovery. The preliminary final evaluator specifies an 80-point minimum-MSE scale grid. The expanded W+A static evaluations and additional architectures use 64 points. These evaluator variants must be distinguished from the 64-point candidate-error profiling grid. We distinguish these variants when interpreting the preliminary ablations and expanded comparisons.

To account for differences in layer sensitivity, we accumulate the squared weight gradients over calibration batches. Let

$$S_n^{W,(t)} = \sum_{i \in W_n} \left(\frac{\partial \mathcal{L}^{(t)}}{\partial w_i} \right)^2 \tag{11}$$

denote the gradient-energy score for layer n at calibration batch t.

The implementation aggregates this quantity using an exponential moving average with $\eta = 0.9$:

$$\widetilde{S}_n^{W,(t)} = \eta \widetilde{S}_n^{W,(t-1)} + (1-\eta) S_n^{W,(t)}. \tag{12}$$

The EMA is initialized with the first batch score. After processing the calibration batches, the square-root transformation and global normalization are applied:

$$\mathcal{F}_n^W = \frac{\sqrt{\widetilde{S}_n^W}}{\max_j \sqrt{\widetilde{S}_j^W}} + 10^{-6}. \tag{13}$$

We refer to $\mathcal{F}_n^W$ as the *normalized gradient-based layer sensitivity score*. This quantity is Fisher-inspired but is not identified as the literal Fisher Information trace.

The sensitivity-weighted weight damage for candidate bit-width b is

$$L_{n,b}^W = \mathcal{F}_n^W \epsilon_{n,b}^W. \tag{14}$$

The aggregate weight contribution is

$$L_W = \sum_n \sum_{b \in \mathcal{C}^W} L_{n,b}^W q_{n,b}^W. \tag{15}$$

A.3 Empirical Global Activation Sensitivity

The global activation profiling procedure perturbs the output of one collected convolutional layer at a time using a temporary symmetric 4-bit quantizer.

For activation tensor z_n, let

$$q_{\max} = 2^{b_{\text{test}}-1} - 1, \qquad b_{\text{test}} = 4, \tag{16}$$

and

$$s_n = \frac{\max |z_n|}{q_{\max}}. \tag{17}$$

The perturbed activation is

$$\widetilde{z}_n = \text{clip}\left(\text{round}\left(\frac{z_n}{s_n} \right), -q_{\max}, q_{\max} \right) s_n. \tag{18}$$

For every layer, the modified network is evaluated on a validation subset. The global activation damage score is the average final reconstruction MSE:

$$G_n^A = \frac{1}{|\mathcal{D}_p|} \sum_{(x,y) \in \mathcal{D}_p} \text{MSE}\left(f_{n,4}(x), y \right), \tag{19}$$

where $f_{n,4}$ denotes the model with the temporary 4-bit output quantization inserted at layer n.

The implementation applies a square-root transformation and normalizes the scores:

$$\mathcal{F}_n^A = \frac{\sqrt{G_n^A}}{\max_j \sqrt{G_j^A}} + 10^{-6}. \tag{20}$$

This ground-truth-loss definition is used in the preliminary all-convolution studies. In the additional architecture experiments, the global probe instead measures disturbance relative to the unperturbed FP32 output: mean squared output distance for restoration and KL divergence from the FP32 class probabilities for classification. The same square-root/max normalization is applied. These are distinct sensitivity variants, not interchangeable loss measurements. The output probe and input-activation MSE are associated with the same convolution, but perturb different tensor locations.

A.4 Candidate-Specific Activation Quantization Error

Candidate-specific activation error is measured separately from the global activation sensitivity. Profiling captures the input activation to each collected convolutional layer. For each candidate bit-width a, the activation is quantized using the same symmetric scale-search procedure described above, and its minimum MSE is recorded.

Let $z_n(x)$ denote the sampled input activation to layer n. The candidate-specific activation error is

$$\epsilon_{n,a}^A = \frac{1}{|\mathcal{D}_a|} \sum_{x \in \mathcal{D}_a} \min_{s \in \mathcal{S}_{n,a}} \mathrm{MSE}\left(Q_a(z_n(x); s), z_n(x)\right). \tag{21}$$

To control memory consumption, the implementation evaluates at most 50,000 activation values for each layer using deterministic strided subsampling.

The activation damage coefficient is therefore

$$L_{n,a}^A = \mathcal{F}_n^A \epsilon_{n,a}^A. \tag{22}$$

The aggregate activation contribution is

$$L_A = \sum_n \sum_{a \in \mathcal{C}^A} L_{n,a}^A q_{n,a}^A. \tag{23}$$

A.5 Asymmetric Activation Search Space

Preliminary joint weight–activation experiments initially allowed lower activation precisions. These configurations resulted in severe degradation, motivating a restricted activation search space for the final formulation:

$$\mathcal{C}^W = \{4, 5, 6, 7, 8\}, \qquad \mathcal{C}^A = \{6, 7, 8\}. \tag{24}$$

The preliminary local-gradient experiments motivated changing both the activation profile and candidate menu (Appendix D.2). They are not a controlled menu-only ablation of this empirical formulation.

A.6 Logarithmic Penalty Sweeping

Because stochastic QUBO optimization can produce non-monotonic changes in the selected architecture as the compute penalty changes, the main Pareto analysis samples the global penalty coefficient γ over a logarithmic range.

For $[\gamma_{\min}, \gamma_{\max}]$, the sampled values are

$$\gamma_i = 10^{\log_{10}(\gamma_{\min}) + i \frac{\log_{10}(\gamma_{\max}) - \log_{10}(\gamma_{\min})}{N-1}}, \qquad i \in \{0, \ldots, N-1\}. \tag{25}$$

Development exploratory sweeps use 500 simulated-annealing reads and the preliminary refinement examples use 1,000 reads for their seed. The expanded target-BOP comparisons use a different procedure: gamma search with five reads per step, at most 20 expansion steps and 14 binary-search steps, followed by an independent 500-read solve. Requested target BOPs therefore need not equal the achieved discrete cost.

B Structural Priors and Solver Details

The QUBO objective incorporates interactions between variables associated with different parts of the computational graph. These terms are implemented as quadratic off-diagonal contributions.

B.1 BOPs-Style Weight–Activation Coupling

Let the summation index range over the selected, routed convolutions. Fixed full-precision layers and non-convolutional operators are excluded from this proxy. The computational cost of a convolutional layer is modeled using the BOPs-style proxy

$$C_{\text{BOPs}} = \frac{\sum_n MAC_n b_n^W b_n^A}{\sum_m MAC_m 32^2}. \tag{26}$$

After substituting the One-Hot representation,

$$C_{\text{BOPs}} = \sum_n \sum_{b \in \mathcal{C}^W} \sum_{a \in \mathcal{C}^A} \left(\frac{MAC_n ba}{\sum_m MAC_m 32^2} \right) q_{n,b}^W q_{n,a}^A. \tag{27}$$

The resulting interaction is multiplied by the global compute penalty γ in the QUBO matrix.

B.2 Skip-Connection Compatibility

The implemented model contains the three-branch fusion

$$X_{\text{fused}} = x_1 + u_2 + u_3. \tag{28}$$

The corresponding activation variables are grouped as

$$\mathcal{A}_{\text{skip}} = \{n_{\text{res}}, n_{\text{up},2}, n_{\text{up},3}\}. \tag{29}$$

Here n_{res} identifies a residual-branch convolution and $n_{\text{up},2}$ and $n_{\text{up},3}$ identify two upsampling-path convolutions selected as fusion-path proxies. For two activation states with different bit-widths, $a_1 \neq a_2$, the assembled QUBO receives an exclusion penalty:

$$Q_{(i,a_1),(j,a_2)} \leftarrow Q_{(i,a_1),(j,a_2)} + 2\alpha. \tag{30}$$

This encourages compatible activation bit-widths among the participating branches. Equal bit-widths do not by themselves imply identical scales, offsets, or dynamic ranges; the constraint is therefore interpreted as a structural compatibility prior rather than a guarantee of identical dynamic ranges or an integer-only addition interface. The activation variables refer to convolution inputs during evaluation; these positions proxy the fusion paths rather than quantizing all three tensors entering the addition.

B.3 Module-Order Activation Interactions

The cross-layer interaction implemented by the proposed QUBO acts on activation variables. The implemented pairs are consecutive convolutions in module traversal order, which is not a traced data-flow graph and can cross parallel paths. For a selected ordered pair n and $n+1$, the interaction between candidate activation states a_i and a_j is

$$Q_{(n,a_i),(n+1,a_j)} \leftarrow Q_{(n,a_i),(n+1,a_j)} + \omega L^A_{n,a_i} L^A_{n+1,a_j}, \tag{31}$$

where

$$L^A_{n,a} = \mathcal{F}^A_n \epsilon^A_{n,a}. \tag{32}$$

The reported advanced formulation uses

$$\omega = 0.15. \tag{33}$$

This interaction penalizes simultaneous large estimated activation damage at paired positions in that ordering. We interpret it as an ordered-layer regularizer, rather than an exact model of error propagation along graph edges.

B.4 Unified QUBO Matrix

Before assembling the QUBO matrix, the diagonal loss coefficients are normalized by their maximum value:

$$L_{\max} = \max_u L_u, \tag{34}$$

$$\widehat{L}_u = \frac{L_u}{L_{\max}}. \tag{35}$$

For adjacent activation states u and v, component construction first stores $L^{\text{cross}}_{uv} = \omega L^A_u L^A_v$; thus ω is already inside the cross coefficient before matrix assembly. Because this term is a product of two loss coefficients, the implementation normalizes it by

$$\widehat{L}^{\text{cross}}_{ij} = \frac{L^{\text{cross}}_{ij}}{L^2_{\max}}. \tag{36}$$

The largest BOP coupling coefficient is denoted by

$$C^{\text{BOP}}_{\max} = \max_{(i,j)} C^{\text{BOP}}_{ij}, \tag{37}$$

and the largest normalized cross-layer loss coefficient by

$$L^{\text{cross}}_{\max} = \max_{(i,j)} \widehat{L}^{\text{cross}}_{ij}. \tag{38}$$

For $\beta = 1$ in the reported implementation, the exclusion penalty is adapted as

$$\alpha = 2\left(\beta + \beta L^{\text{cross}}_{\max} + \gamma C^{\text{BOP}}_{\max} + 1\right). \tag{39}$$

Let $x_u \in \{0, 1\}$ represent a precision-assignment state. The QUBO energy is

$$E(x) = \sum_u Q_{u,u} x_u + \sum_{u<v} Q_{u,v} x_u x_v. \tag{40}$$

The implemented diagonal coefficients are

$$Q_{u,u} = \beta \widehat{L}_u - \alpha. \tag{41}$$

For conflicting one-hot states,

$$Q_{u,v} \mathrel{+}= 2\alpha. \tag{42}$$

For weight–activation BOP coupling,

$$Q_{u,v} \mathrel{+}= \gamma C_{u,v}^{\mathrm{BOP}}. \tag{43}$$

For adjacent-activation interactions,

$$Q_{u,v} \mathrel{+}= \beta \widehat{L}_{u,v}^{\mathrm{cross}}. \tag{44}$$

Here $\widehat{L}_{u,v}^{\mathrm{cross}}$ already contains the factor ω; $\beta = 1$ is the reported assembly-scale multiplier.

The skip-mismatch pairs also receive 2α, as described above; multiple contributions to the same pair are added. All remaining entries are zero.

This explicit matrix construction corresponds to the numerical QUBO assembled in the implementation.

B.5 Simulated Annealing Search

The QUBO matrix is optimized using classical simulated annealing. For the primary experiments, the solver evaluates the requested number of annealing reads with random seed 123. The seed is exposed explicitly so that stochastic robustness can be assessed in a separate multi-seed experiment. For each run, the configuration with the lowest observed QUBO energy is selected:

$$x^* = \arg \min_{x \in \mathcal{X}_{\mathrm{observed}}} E(x). \tag{45}$$

Because Simulated Annealing is a stochastic heuristic and the search is not exhaustive, x^* is treated as the *best observed solution under the specified solver budget*, rather than as a certified global optimum. The extended solver then repairs invalid one-hot groups to their highest precision state and performs up to four coordinate-improvement passes over layer groups. This postprocessing is separate from raw annealing; reported raw solver energy need not equal the energy of a subsequently repaired route. The corrected all-convolution seed study uses its separate raw-solver path.

C Architecture and PROTES Implementation

The QUBO stage evaluates candidate precision assignments using a surrogate objective combining sensitivity-weighted quantization damage and computational cost. PROTES is subsequently used to refine the best observed QUBO solution using direct network evaluation.

C.1 Model Architecture

The experimental HalfUNet model is a custom compact encoder–decoder. The architecture contains NAFBlock-based residual units, depthwise convolution, pointwise convolution, channel-wise layer normalization, multiplicative gating, simplified channel attention, and pixel-shuffle upsampling. Its final representation is obtained using the three-way feature fusion described previously.

The NAFBlock design is based on NAFNet Chen et al. (2022). However, the overall network used in this work is a custom architecture and is not the canonical NAFNet or the published Half-UNet architecture.

C.2 PROTES Search Space and QUBO Seeding

For a model containing L optimized layers, PROTES operates on a discrete index vector of dimension

$$d = 2L, \tag{46}$$

where the first L coordinates represent weight precision states and the remaining L coordinates represent activation precision states.

The best observed QUBO solution is converted to the corresponding PROTES index vector $i_\star$. This vector is used to construct a peaked initial TT probability distribution. The same solution is also supplied through the seed-pool mechanism during the first five PROTES iterations.

The resulting procedure is a seeded probabilistic search; it does not impose a hard Hamming-distance constraint around the QUBO solution.

C.3 ROI-Constrained PROTES Objective

Let B be the candidate's normalized routed-layer BOP ratio and $B_0 = B(\mathcal{A}_{\mathrm{QUBO}})$ the seed ratio. The implemented restoration objective is the continuous piecewise score

$$S(\mathcal{A}) = \mathrm{PSNR}(\mathcal{A}) - \gamma B - (h - \gamma)\max(B - B_0, 0), \qquad h = \max(10\gamma, 100). \tag{47}$$

Thus, for $B \le B_0$ the cost is γB, while above the seed it is $\gamma B_0 + h(B - B_0)$. Only the excess receives the larger slope; there is no discontinuous charge on the whole candidate cost. At $\gamma = 1.584893$, increasing B by 0.01 above the seed incurs an additional 1.0 dB penalty. This is a soft penalty, not a hard compute constraint, so selected candidates can exceed the seed cost. The QUBO coefficient and the black-box penalty are reused numerically in these runs, but operate on different objective scales. For classification, top-1 accuracy in percentage points replaces PSNR in the direct-evaluation score.

C.4 PROTES Evaluation Strategy

For the repeated-search experiment, PROTES uses sampling batch size 50, ten elites, tensor-train rank five and learning rate 0.05. The seeded initialization uses concentration eight and exploration probability 0.03; 30% seed-pool sampling is enabled for five iterations with five random samples. Search seeds are 1701, 1702 and 1703, and recovery uses seed 77003 for the selected target. The five validation-subset runs keep optimizer seed 4402 fixed, with subset-index seed 4401 and recovery seed 88003.

The optimizer seed of the preliminary W+A sweep is unverified. The repeated-search experiment explicitly controls the optimizer seeds listed above; these studies therefore provide different evidence about stochastic variation.

During optimization, each candidate architecture is evaluated using a 32-patch validation subset to reduce the cost of repeated network evaluation. The best candidate found by PROTES is subsequently evaluated on the complete test loader.

The reported PROTES configuration uses a batch size of 50 candidates, keeps the 10 highest-scoring candidates as elites, uses TT rank 5, an Adam learning rate of 5×10^{-2}, and peaked initialization around the QUBO solution, a seed-pool fraction of 0.3 during the first five iterations, and an exploration probability of 0.03. The primary full-sweep cost records report a nominal budget of 5,000; the matched-method and subset-robustness experiments use 2,000. Development refinement examples report their own 1,000/2,000 budgets. The primary search pads the activation alphabet to match the five weight choices and copies the first activation choice in each coupled group to the others; thus different sampled indices may decode to the same route. Cache counts in the original sweep count sampled index vectors, whereas the repeated- search study counts canonical assignments.

Table 4: Experiment variants and evidence scope. Costs use the routed-layer FP32 denominator; static and recovered results are reported separately.

Record	Routing and split	Scope
Primary comparison	38 convolutions / 304 variables; patch-random SIDD	Single original recovery evaluations; repeated routes at loose targets
Repeated search and subset studies	Protected primary routing; 32-patch validation inner loop	Search-seed/subset variation conditional on fixed recovery settings
Corrected paired statistics and solver robustness	40 convolutions / 320 variables; patch-random SIDD	Static paired patch tests and solver-seed variation
Supporting ablations	All-convolution variant	Exploratory supporting ablations; not recovered main-result significance
Public pretrained SIDD	224 / 260 / 83 routed convolutions; official 40-image validation blocks	Five calibration images and 35 evaluation images; static fake quantization only
QUBO solver benchmark	Four fixed problems with 1,792 / 2,080 / 664 / 304 variables	Five trials for each of 19 backends; common projection and network evaluation
PROTES statistical extension	40 convolutions / 320 variables; patch-random SIDD	Paired patch tests, three refinement seeds, objective and activation-menu sweeps
Formulation development	Earlier routing, menus and optimizer budgets	Contextual observations; not pooled with the primary comparison

D Full Experimental Record

We evaluate the framework through a sequence of experiments that isolate the effects of task-aware sensitivity, activation handling, topological constraints, and PROTES refinement.

D.1 Implementation Details and Metrics

The repeated-search runs use Linux, Python 3.10.20, PyTorch 2.9.1+cu128, CUDA 12.8 and an NVIDIA H100 80GB HBM3 GPU; the repeated-search environment uses cuDNN 9.2. The public-checkpoint SIDD and 19-solver runs instead use a consumer laptop with an RTX 3060 Laptop GPU and Ryzen 7 6800H CPU. Solver times are therefore laptop-specific. The PROTES statistical extension ran on a separate server whose specifications were not retained; we report its timing only within that study. None of these settings are retroactively assigned to preliminary runs. Repeated-search seeds are 1701–1703; its recovery seed is $77000 + t$, where t is the zero-based target index. The subset study uses seeds 4401 (selection), 4402 (search) and 88000 (recovery). Additional-architecture recovery uses base seed 42 with target offset $1000t$.

The primary architecture is the custom NAFBlock-based HalfUNet implementation described above. It accepts three-channel inputs and uses a base channel width of 32. The network contains two encoder downsampling stages, a central NAFBlock stage, two pixel-shuffle upsampling paths, and a three-branch feature-fusion operation.

The primary experiments use 160 noisy/ground-truth pairs from SIDD Small sRGB Abdelhamed et al. (2018). Each image is divided into four quadrants, which are resized to 256×256 RGB patches. The resulting 640 pairs are randomly split into 512 training, 64 validation and 64 test patches using seed 42. Each training patch is accessed three times, with an unaugmented access and stochastic flip augmentations, giving 1,536 training items. Validation and test patches are not augmented; the primary batch size is 16.

This split occurs after patch extraction. Different patches from the same source image can therefore appear in different partitions. We call it the *patch-random* protocol; these resized-patch scores are not official full-resolution SIDD benchmark results.

The expanded mixed-precision comparison protects the first and last convolutions, routing 38 internal convolutions (304 binary variables). The paired statistics, corrected SA-seed study, omega/beta, probe-bit, activation-menu and bias studies instead route all 40 convolutions (320 variables). We report these as an *all-convolution variant*, not as repetitions of the protected-layer main comparison.

For final architecture evaluation, the baseline network is deep-copied and the selected weight bit-width is applied independently to each selected convolutional layer. Development weight quantization uses the symmetric minimum-MSE scale search described above with an 80-point scale grid; expanded evaluations use 64 points as specified in Appendix A. Convolutional biases are not quantized in the reported experiments and therefore remain at their original precision. Activation quantization is applied dynamically during forward evaluation using the selected activation bit-width.

The computational cost is represented by a normalized BOPs proxy. For layer n,

$$BOPs_n = MACs_n b_{W,n} b_{A,n}. \tag{48}$$

The Compute Cost Ratio is

$$\text{Compute Cost Ratio} = \frac{\sum_n MACs_n b_{W,n} b_{A,n}}{\sum_n MACs_n 32^2}. \tag{49}$$

This quantity is a computational proxy based on the assigned precision and MAC count; it is not intended to represent measured hardware latency.

Network restoration quality is evaluated using PSNR and SSIM Wang et al. (2004), with RGB outputs clipped to $[0, 1]$ and data range one. The shared restoration evaluator computes PSNR from the MSE over each batch and averages the batch scores. The paired analysis separately averages per-patch PSNR, which is a different aggregation. Dynamic symmetric activation quantization uses the maximum magnitude over the current tensor, including its batch dimension; batch size is therefore part of the protocol. Average weight bits are parameter-count weighted, average activation bits are MAC weighted, and selected-weight storage reduction is $1 - \sum_n |W_n| b_n^W / (32 \sum_n |W_n|)$. Development layer-average bit counts are unweighted unless otherwise specified.

The static implementation retains floating-point weights after rounding and uses floating-point convolution, bias, accumulation, normalization, residual addition and gating arithmetic. The extended LSQ+ protocol uses unsigned asymmetric activation quantizers, four epochs updating quantizer parameters and fourteen epochs also updating convolution parameters, with a cap of 100 batches per epoch. Adam step sizes are 10^{-4} for scales/offsets and 10^{-5} for convolution parameters; scale/offset rates are halved in the second stage. Restoration recovery uses L1 loss; classification uses cross-entropy. This recovery is distinct from the MSE-based preliminary sensitivity calculation and from isolated scale calibration.

The profiling caps are 10 validation batches for weight gradients and candidate activation error and five for the global activation probe. With 64 validation patches at batch size 16, the primary loader contains only four batches, so these caps do not imply ten distinct processed batches. All sensitivity/calibration data come from validation, not the test split.

D.2 Formulation Development: Full Method Chronology

The preliminary development records explain two distinct revisions: the weight-damage objective and the activation-sensitivity estimator. They are before-recovery evaluations from earlier protocols, not controlled ablations of the final protected-layer comparison.

Weight-only development. The original bit-reduction objective produced an imbalanced route: 35.778 dB at a reported average of 7.94 weight bits. One-hot selection using parameter-count-weighted reconstruction MSE reached 36.803 dB near six bits. Replacing that damage proxy with normalized squared-gradient sensitivity and removing the additional parameter-count multiplier reached 36.925 dB (Table 5). The 0.122 dB difference between the two one-hot routes is the recorded incremental gain of task-aware weighting; relative to uniform six-bit weights, the gain is 0.298 dB. These preliminary averages do not establish measured storage savings or solver optimality.

Joint-allocation phases. At $\gamma = 20$, Phase 1 reused the weight-gradient score for both weights and activations. Phase 2, also reported at $\gamma = 20$, instead collected squared activation gradients

during backpropagation and used separate local scores for the two tensors. It yielded 25.481 dB at reported averages of 6.10 weight bits and 6.00 activation bits. Phase 3 multiplied the activation-damage term by $\lambda_A = 50$; its $\gamma = 5$ run recorded 24.896 dB shows that scalar reweighting did not resolve the failure at that operating point. A temporary fallback retained QUBO weight routing and fixed activations at eight bits, reaching 36.448 dB. These are different estimators and objectives, not simply an unrestricted-activation variant of the final empirical method.

The later method perturbed one intermediate activation with temporary four-bit quantization and measured the resulting reconstruction MSE at the network output across validation data. It used that downstream response to construct a global empirical activation-sensitivity score, alongside candidate activation reconstruction errors. It also restricted the activation menu to $\{6, 7, 8\}$ while allowing weights in $\{4, 5, 6, 7, 8\}$. Its reported $\gamma = 50$ route used 5.58-bit weights and 7.17-bit activations and achieved 35.105 dB at 3.603% BOPs (Table 6). This activation average is consistent with the final menu. Profiling and the menu changed together, so the preliminary phase comparison cannot isolate their separate contributions or prove an intrinsic six-bit activation threshold. The later menu-only study appears separately in Appendix D.16. This is the same no-structural-priors route reported in Table 2, not an additional experiment.

Table 5: Weight-only objective-development comparisons before LSQ+. Activations are not jointly allocated; average precision and PSNR use the preliminary evaluation protocol summarized in Appendix D.2.

Weight allocation	Avg. W bits	PSNR (dB)
FP32 reference	32.00	37.263
Uniform eight-bit	8.00	37.207
Original bit-reduction QUBO	7.94	35.778
Uniform six-bit	6.00	36.627
One-hot reconstruction QUBO	6.00	36.803
One-hot task-aware QUBO	6.03	36.925

Table 6: Joint W+A formulation-development comparisons before LSQ+. Average bits are layer averages; approximate BOPs are report estimates. Profiling and menus vary across these preliminary settings (Appendix D.2). The fixed-A8 route's 4.63% entry is estimated from the product of its layer-average weight and activation bits, not a recorded MAC-weighted cost.

Strategy	Avg. W	Avg. A	BOPs (%)	PSNR	SSIM
Uniform W4/A8	4.00	8.00	3.125	33.405	0.886
Uniform W6/A8	6.00	8.00	4.688	36.297	0.917
Shared weight-gradient score	6.15	6.15	3.75	33.750	0.866
Separate local W/A gradients	6.10	6.00	3.60	25.481	0.447
Local gradients + A scalar	5.62	6.50	3.55	24.896	0.431
QUBO W + fixed A8	5.92	8.00	4.63	36.448	0.917
Empirical + asymmetric menu (no priors)	5.58	7.17	3.603	35.105	0.887

Table 7: Refinement objective-development comparison before LSQ+, with a nominal budget of 2,000 per run. Cells give achieved BOPs (%) / PSNR (dB), starting from the displayed seed. These preliminary records use the development protocol, separately from Figure 4.

γ	QUBO seed	Linear-score refinement	Expansion-penalized refinement
85.77	3.323 / 35.144	3.113 / 35.573	3.165 / 35.704
11.66	4.119 / 35.969	4.790 / 36.586	4.042 / 36.300
1.58	3.990 / 35.497	4.989 / 36.460	3.985 / 36.318

D.3 Ablation Study: Structural Priors

We compare the explored quality–cost operating points for three formulations:

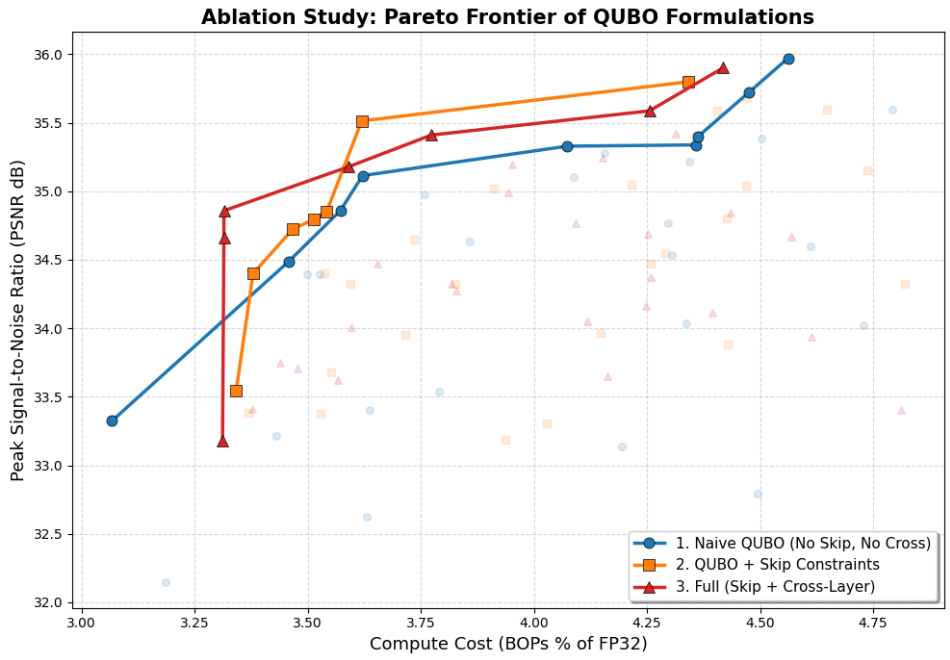


Figure 5: Ablation study of QUBO formulations. Scattered points represent candidate architectures evaluated during the stochastic search, while the solid curves connect reported operating points and are not a certified Pareto frontier. Adding skip-connection and adjacent-activation interactions improves the observed PSNR–compute tradeoff below approximately 3.6% BOPs; the curves cross at looser operating points.

1. **Naive QUBO (No Skip, No Cross):** the task-aware formulation without graph-aware interactions.
2. **QUBO + Skip Constraints:** the task-aware formulation with the skip-connection compatibility penalties.
3. **Full (Skip + Cross-Layer):** the formulation additionally including the adjacent-activation interaction with $\omega = 0.15$.

The Naive QUBO formulation exhibits pronounced performance degradation below approximately 3.5% BOPs. The Skip-Constrained formulation improves the observed frontier in the intermediate compression regime, while the Full formulation provides additional recovery at more aggressive compression levels.

The Full formulation reaches approximately 35.9 dB at 4.4% BOPs in the reported sweep. These results support the use of selected structural priors within the evaluated QUBO formulation, but they do not establish global optimality over the complete search space.

D.4 Development Study: QUBO-Seeded PROTES

The best observed QUBO configuration is used to initialize PROTES. For the reported end-to-end refinement experiment, the QUBO seed is obtained with 1000 Simulated Annealing reads.

The report describes an expansion tax with rate $\max(10\gamma, 100)$ and evaluates refinement on 32 validation patches. These preliminary records are separate from the expanded matched-method comparison; the report description does not establish exact equivalence to the continuous refinement score in Eq. 47. The linear-score comparison appears in main Table 7.

The comparison at $\gamma = 85.77$ uses two different PROTES evaluation budgets, $m = 1000$ and $m = 2000$, while keeping the QUBO seed fixed. These are single stochastic runs with different budgets, not a repeated controlled estimate of the budget effect.

Table 8: ROI-Constrained PROTES Refinement across Selected γ Regimes.

Target γ	Budget (m)	QUBO Seed	PROTES Final	Δ BOPs	Δ PSNR
$\gamma = 85.77$	1000	3.323% / 35.144 dB	3.308% / 35.693 dB	-0.016%	+0.549 dB
$\gamma = 85.77$	2000	3.323% / 35.144 dB	3.165% / 35.704 dB	-0.158%	+0.560 dB
$\gamma = 11.66$	2000	4.119% / 35.969 dB	4.042% / 36.300 dB	-0.078%	+0.331 dB
$\gamma = 1.58$	2000	3.990% / 35.497 dB	3.985% / 36.318 dB	-0.004%	+0.821 dB

Table 9: Representative computational and storage characteristics at the executed deployment operating point. BOPs and ideal packed storage are analytical quantities; the current reference implementation does not pack weights into arbitrary-bit tensors.

Method	BOPs (%)	Whole-model reduction (%)	Ideal packed size (KiB)
Uniform target-matched	4.102	77.592	61.31
HAWQ-style Hessian + knapsack	4.101	78.486	58.87
Enhanced QUBO	4.149	79.745	55.43
Enhanced QUBO + PROTES	4.035	79.080	57.25

D.5 Main Mixed-Precision Comparison

The HAWQ-style baseline combines normalized absolute Hutchinson weight-Hessian trace estimates with calibrated weight reconstruction error and the shared activation-error/sensitivity term. It uses two one-sample batches and two probes per batch (seed 2026), followed by a multiple-choice dynamic-programming allocator with resource resolution 100,000. This is a sensitivity–allocation adaptation, not the complete HAWQ-V3 deployment system.

Main Table 1 reports the original six-target records, reporting achieved rather than requested cost. Static/LSQ+ rows are paired by assignment ID; reused routes share their static evaluation. Later additional uniform recovery runs and dense gamma-sweep points are not pooled into these single-run rows. These are descriptive operating points, not seed-averaged estimates or equal-cost comparisons.

At the 4.102% requested target, QUBO + PROTES achieves 37.192 dB after LSQ+ at 4.035% BOPs, compared with 37.132 dB at 4.149% for QUBO and 37.092 dB at 4.101% for HAWQ-style allocation. The original uniform row gives 36.802 dB; an additional recovery run on that same uniform assignment gives 37.010 dB, illustrating variation not captured by a single row. At the lowest target, QUBO alone reaches 36.997 dB after LSQ+, exceeding the refined route's 36.823 dB despite the latter's stronger static result. Refinement gains depend on both the operating point and recovery.

The three loosest requested targets reuse the same QUBO assignment and the same refined assignment within each method. Their differing LSQ+ scores reflect separate recovery runs, not three distinct routing solutions. No EdMIPS or ALPS/EAGL comparison is included in the completed experiments.

Weight-storage compression is reported numerically at the representative operating point used for the deployment analysis. Because the current implementation keeps floating-point tensors in the serialized model state, we distinguish the analytical packed-storage reduction from the realized serialized model size.

D.6 Public Pretrained SIDD Benchmark

We evaluate public NAFNet-width32, its activation-bearing Baseline-width32, and HINet-0.5x checkpoints Chen et al. (2022; 2021). The official SIDD validation tensor contains 40 source images with 32 blocks each. Images 0, 8, 16, 24 and 32 (160 blocks) are used only for profiling and search; the remaining 35 images (1,120 blocks) are used only for final evaluation. The first and last Conv2d layers remain FP32, as do HINet ConvTranspose2d layers, leaving 224, 260 and 83 routed convolutions, respectively. All methods target $42/1024 = 4.1015625\%$ routed-layer BOPs. PROTES uses 2,000 proposals and 32 uniformly spaced calibration blocks. No dense training or LSQ+ recovery is performed in this benchmark.

Table 10: Training-free static results using public pretrained SIDD checkpoints. Costs are routed-layer BOP percentages relative to FP32.

Model	Method	BOPs (%)	PSNR (dB)	SSIM
NAFNet-width32	FP32	100.000	39.9047	0.92135
	Uniform W6/A7	4.1016	28.1094	0.61619
	HAWQ-style	4.0783	38.1711	0.89332
	QUBO	4.1005	38.5028	0.90381
	QUBO + PROTES	4.0869	**39.5153**	**0.91685**
Baseline-width32	FP32	100.000	39.8151	0.92074
	Uniform W6/A7	4.1016	38.5404	0.90828
	HAWQ-style	4.1064	39.0306	0.90969
	QUBO	4.1035	39.1445	0.91279
	QUBO + PROTES	3.9994	**39.4166**	**0.91674**
HINet-0.5x	FP32	100.000	39.7600	0.92057
	Uniform W6/A7	4.1016	37.5512	0.88488
	HAWQ-style	4.0943	38.4354	0.90086
	QUBO	4.0976	38.3518	0.89930
	QUBO + PROTES	4.0484	**38.8902**	**0.90820**

The recorded PROTES search times are 24,579 s for NAFNet, 26,858 s for the Baseline and 13,996 s for HINet on the evaluation laptop. These end-to-end search times include repeated network evaluation and should not be compared with the solve-only QUBO times below.

D.7 QUBO Solver Benchmark

We hold each model's QUBO fixed and compare 19 backends over five trials. The problem sizes are 1,792 variables for NAFNet, 2,080 for the Baseline, 664 for HINet and 304 for HalfUNet. Simulated annealing uses 16 reads and 2,000 sweeps; the dynamical methods use 64 parallel states and 1,000 steps. The QAIA rows execute the original MindQuantum source Xu et al. (2024) on the laptop GPU. Each raw output is projected to a valid one-hot assignment before energy and BOP evaluation. Table 11 ranks each backend's best projected-energy trial within each model. Ties are retained. This standardized backend test is intentionally separate from the production SA pipeline, which used gamma search, an independent 500-read final solve and coordinate improvement. Consequently, standardized SA is a controlled budget reference rather than a rerun of the production method.

Table 11: All 19 QUBO backends across four model problems, listed one per row. "Budget" counts models whose best-energy assignment is within 0.25 percentage points of the 4.102% target. Time is the median solve-only time across the four selected best trials on the RTX 3060 laptop.

Solver	Mean rank	Wins	Budget	Time (s)
SFC	1.25	3	4/4	1.23
LQA	1.50	2	4/4	1.50
NMFA	3.00	0	1/4	1.19
SimCIM	4.25	0	2/4	1.36
CAC	5.25	0	4/4	1.16
Tabu	6.00	0	0/4	320.94
BSB	7.25	0	3/4	0.75
ASB	8.25	0	4/4	1.08
Steepest descent	8.50	0	4/4	0.03
GBSB	10.25	0	3/4	0.93
USB	12.00	0	3/4	0.98
TSB	12.50	0	3/4	0.91
LSB	13.25	0	2/4	1.55
Rotor SQA	14.00	0	3/4	1.63
CFC	14.75	0	3/4	1.30
Simulated annealing	15.00	0	4/4	0.18
Path-integral SQA	16.00	0	4/4	28.79
DSB	18.00	0	2/4	0.78
PulseCIM	18.75	0	3/4	0.88

For network evaluation, each backend contributes the trial closest to the target BOP ratio, with projected energy breaking ties. This differs from the best-energy ranking above and prevents large budget deviations from being rewarded solely for lower surrogate energy. Tables 12 and 13 report the evaluation of all 19 selected assignments; the earlier compact table reported only representative rows. For the three public checkpoints, production SA/QUBO records 38.503, 39.145 and 38.352 dB; the corresponding SFC values are 38.502, 39.135 and 38.352 dB. The HalfUNet column uses batch-one evaluation, whereas the primary benchmark uses batch size 16. Because activation scales include the batch dimension and PSNR is aggregated per batch, HalfUNet values in this subsection are internally comparable across solvers but not to the primary FP32 or method scores.

Table 12: Network evaluation for all 19 solver backends. Each entry is PSNR (dB) / achieved BOPs (%). Rows follow the projected-energy ranking in Table 11; quality comparisons should account for budget mismatch. HalfUNet uses a separate batch-one evaluation protocol and must not be compared with the batch-16 primary table.

Solver	NAFNet-width32	Baseline-width32	HINet-0.5x	HalfUNet (batch 1)
SFC	38.502/4.091	39.135/4.057	38.352/4.091	37.700/4.220
LQA	38.502/4.095	39.140/4.079	38.353/4.089	37.700/4.220
NMFA	38.618/4.660	39.277/4.646	38.480/4.204	37.810/4.498
SimCIM	38.577/4.552	39.247/4.552	38.008/4.087	37.377/4.123
CAC	38.465/4.100	38.818/4.000	38.351/4.124	37.258/4.129
Tabu	38.556/5.181	39.196/5.098	38.165/4.940	37.574/4.209
BSB	25.069/4.036	39.101/4.322	38.003/4.124	37.074/4.058
ASB	38.498/4.160	39.006/4.088	37.831/3.952	36.983/4.257
Steepest descent	38.267/4.121	38.440/4.124	37.947/3.960	37.169/4.116
GBSB	24.536/4.082	37.854/4.288	37.520/4.107	37.424/4.090
USB	20.834/4.099	38.799/4.115	37.692/4.029	37.001/4.124
TSB	22.051/4.093	37.253/4.110	37.851/3.934	37.402/4.109
LSB	37.087/3.898	38.136/4.082	35.917/3.636	37.009/4.247
Rotor SQA	25.609/4.099	38.521/4.089	36.250/4.128	37.389/4.124
CFC	38.392/3.916	39.034/3.992	37.022/3.780	37.444/4.056
Simulated annealing	35.515/4.101	37.456/4.114	36.373/4.108	36.526/4.074
Path-integral SQA	24.265/4.085	38.683/4.141	36.388/4.190	36.909/4.073
DSB	38.279/4.069	38.181/3.825	36.445/3.782	37.061/3.884
PulseCIM	38.361/4.014	38.826/4.046	38.020/4.001	36.828/4.108

Figure 6 compares each backend with the production SA/QUBO route on the three directly compatible public-checkpoint protocols. SFC and LQA reproduce production quality to within 0.01 dB, whereas several other backends fail sharply on NAFNet despite remaining competitive elsewhere. This model dependence reinforces the need to evaluate the selected assignment in the network rather than infer quality from QUBO energy alone.

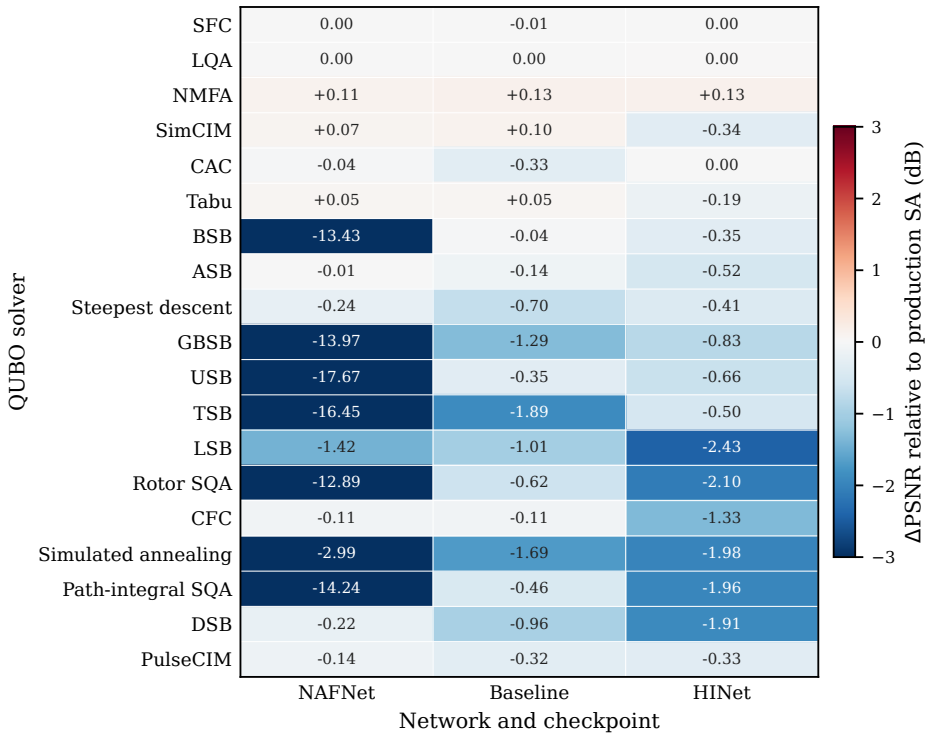


Figure 6: Static PSNR difference relative to the production SA/QUBO route on the three public pretrained checkpoints for the 19 budget-selected solver assignments in Table 12. Numbers are exact differences in dB. To retain contrast among competitive solvers, the color scale saturates outside $[-3, 3]$ dB while annotations show uncapped values.

Table 13: SSIM for the same 19 budget-selected assignments in Table 12. The HalfUNet column again uses batch one.

Solver	NAFNet-width32	Baseline-width32	HINet-0.5x	HalfUNet (batch 1)
SFC	0.9038	0.9128	0.8993	0.9246
LQA	0.9038	0.9128	0.8993	0.9246
NMFA	0.9048	0.9135	0.9019	0.9265
SimCIM	0.9043	0.9133	0.8953	0.9189
CAC	0.9027	0.9107	0.8993	0.9217
Tabu	0.9043	0.9132	0.9008	0.9219
BSB	0.5484	0.9116	0.8950	0.9206
ASB	0.9035	0.9102	0.8937	0.9131
Steepest descent	0.8991	0.8992	0.8909	0.9183
GBSB	0.5310	0.9014	0.8860	0.9222
USB	0.4106	0.9084	0.8824	0.9121
TSB	0.4602	0.8833	0.8899	0.9194
LSB	0.8696	0.8999	0.8478	0.9147
Rotor SQA	0.5645	0.9024	0.8557	0.9192
CFC	0.9013	0.9115	0.8736	0.9207
Simulated annealing	0.8231	0.8968	0.8655	0.9110
Path-integral SQA	0.5297	0.9048	0.8585	0.9179
DSB	0.8998	0.9012	0.8587	0.9109
PulseCIM	0.8998	0.9070	0.8922	0.9133

SFC and LQA produce the lowest or tied-lowest projected energies on all four problems and nearly identical network scores. This agreement does not imply a global optimum. PulseCIM has a median

solve-only time of 0.88 s and a mean energy rank of 18.75. Network evaluation takes approximately 2.4–2.8 s for HalfUNet and 88–222 s for the larger pretrained models, so solver time alone does not determine experiment cost.

D.8 Paired Static Comparison after PROTES Refinement

We evaluate the QUBO+PROTES assignment, its exact QUBO seed and Uniform W4/A8 on the same 64 patches from the held-out patch split with shared source images. The refined assignment is obtained from the saved all-convolution QUBO at $(\omega, \beta, \gamma) = (0.15, 1, 85.77)$. It uses 3.1403% BOPs, below the 3.6962% seed and close to the 3.1250% uniform route. Table 14 reports all three pairwise comparisons from this single saved experiment, keeping the assignments and evaluation protocol internally consistent.

PROTES improves every patch relative to both its QUBO seed and Uniform W4/A8. The seed itself improves 61 of 64 patches relative to uniform. The large effect sizes are specific to this static all-convolution assignment. Patches from related source images and batch-dependent activation scaling limit the independence assumption. We therefore supplement the nominal patch-level tests with a 100,000-replicate cluster bootstrap over the 53 source images represented in the split. This accounts for within-source patch clustering, but it does not remove source-image overlap with the training partition or produce a post-recovery claim.

Table 14: Complete paired static comparison on the same 64 patches from the held-out patch split with shared source images. P, Q and U denote QUBO+PROTES, its exact QUBO seed and Uniform W4/A8, respectively. Assignment values are mean ± patch standard deviation; paired statistics use within-patch PSNR differences. The source-image bootstrap resamples all patches belonging to a sampled source image and uses seed 2026.

Assignment-level results Metric	QUBO+PROTES (P)	QUBO seed (Q)	Uniform W4/A8 (U)
BOPs (%)	3.1403	3.6962	3.1250
PSNR (dB)	36.3751 ± 2.1894	34.9936 ± 1.9583	32.9624 ± 1.3297
Paired differences, $n = 64$ Statistic	P–Q	P–U	Q–U
Mean ΔPSNR (dB)	+1.3815	+3.4127	+2.0311
SD of paired Δ (dB)	0.5878	1.2722	1.0340
95% CI (dB)	[1.2347, 1.5284]	[3.0949, 3.7305]	[1.7729, 2.2894]
Source-image bootstrap 95% CI	[1.2261, 1.5491]	[3.1048, 3.7221]	[1.7715, 2.2914]
Paired t statistic	18.803	21.460	15.715
Paired t-test p	1.57×10^{-27}	1.15×10^{-30}	1.78×10^{-23}
Wilcoxon statistic	0	0	8
Wilcoxon p	3.53×10^{-12}	3.53×10^{-12}	5.15×10^{-12}
Paired Cohen's d_z	2.350	2.683	1.964
First better / second better	64 / 0	64 / 0	61 / 3

Figure 7 shows the complete paired distributions underlying Table 14. The positive support of both distributions confirms that the reported means are not produced by a small number of unusually favorable patches.

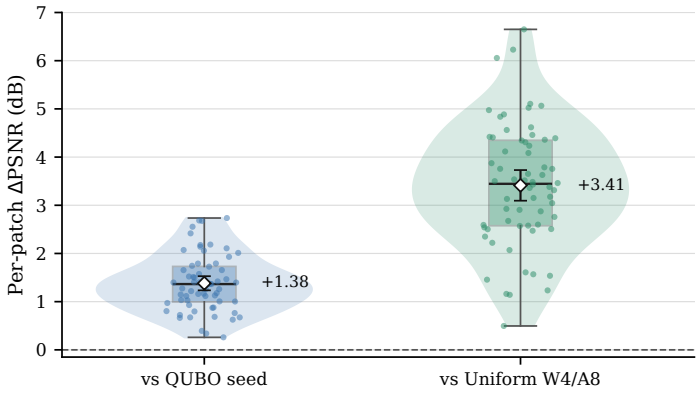


Figure 7: Per-patch static PSNR gains of QUBO+PROTES over its exact QUBO seed and Uniform W4/A8 on the same 64 held-out patches. Points are individual paired differences; violins and boxes summarize their distributions. Diamonds denote means and error bars denote the nominal paired 95% confidence intervals. All 128 plotted differences are positive.

D.9 Search-Method and Initialization Analysis

At a requested BOP ratio of 4.1015625%, we compare four stochastic search procedures using seeds 1701, 1702, and 1703 and a nominal budget of 2,000 proposals per run. All use the same 32-patch validation subset. We also evaluate the fixed QUBO assignment used to initialize refinement. Actual proposal counts, unique network evaluations, and achieved BOP ratios differ between methods and are reported explicitly.

Main Figure 4 visualizes the quality comparison; Table 15 retains the exact scores, costs and evaluation counts.

Table 15: Exact search comparison at one requested compute target. Quality values are means on the held-out patch split with shared source images and, for the stochastic methods, standard deviations over three search seeds. QUBO alone is one fixed assignment; LSQ+ variation is conditional on one fixed recovery seed.

Method	BOPs (%)	Static PSNR	LSQ+ PSNR	Unique evals.
QUBO alone	4.149	36.648	37.153	1
QUBO-seeded PROTES	4.129	36.694 ± 0.081	37.160 ± 0.015	1,953
Random PROTES	4.079	36.550 ± 0.026	37.040 ± 0.082	2,000
Random search	4.094	36.074 ± 0.158	37.020 ± 0.037	2,000
Simulated annealing	4.143	36.617 ± 0.109	37.137 ± 0.019	1,700

Random search uses 2,000 proposals and 2,000 unique evaluations on average. Random PROTES uses 2,001 proposals and 2,000 unique evaluations; QUBO-seeded PROTES uses 2,001 proposals and approximately 1,952.7 unique evaluations. Simulated annealing uses 2,006 proposals and 1,700 unique evaluations. The fixed QUBO assignment is evaluated once. These counts distinguish nominal search budgets from realized black-box evaluation costs.

QUBO-seeded PROTES achieves the highest mean static PSNR among the tested stochastic procedures. Relative to the fixed QUBO route, its mean improvement is approximately 0.046 dB before LSQ+ and 0.007 dB after LSQ+. These are descriptive differences at slightly different achieved BOP ratios. They support a limited benefit of refinement at this operating point, but do not establish general superiority over QUBO alone or the other search procedures.

The mean search runtimes are 1,206.6 s for random search, 1,262.6 s for random PROTES, 1,058.2 s for black-box SA and 1,249.8 s for seeded PROTES. These exclude final LSQ+ recovery. The QUBO-alone timer only evaluates an already selected route; profiling and solving are reported sep-

Table 16: Measured inference latency at the representative operating point. Latency is measured on the reference PyTorch implementation with fake quantization followed by floating-point kernels.

Method	CPU (ms)	GPU (ms)	Speedup vs. FP32 (GPU)
FP32	120.36 ± 67.61	3.002 ± 0.094	1.000
Uniform	188.71 ± 135.86	6.776 ± 0.239	0.443
HAWQ-style	169.85 ± 102.63	6.743 ± 0.196	0.445
QUBO	162.50 ± 77.66	6.735 ± 0.249	0.446
QUBO + PROTES	179.27 ± 100.54	6.769 ± 0.214	0.444

arately in Section D.18. Black-box SA searches on the network score and is distinct from annealing the QUBO surrogate.

D.10 Hardware and Deployment Analysis

Because normalized BOPs are a computational proxy, we additionally measured wall-clock inference latency for representative assignments on the available CPU and GPU hardware. The benchmark uses the same input resolution and model implementation as the primary evaluation. The current reference implementation performs fake quantization and then executes floating-point convolution kernels; it therefore does not provide native arbitrary-bit arithmetic.

This distinction is important when interpreting the BOP results. The BOP ratio and storage-compression values describe the assigned numerical precision, while measured latency reflects the implementation used to execute those assignments. Consequently, a reduction in analytical BOPs is not expected to yield a proportional latency reduction until native low-bit kernels, packed memory layouts, and suitable hardware support are introduced. We therefore do not use the present latency measurements as evidence of hardware acceleration; instead, they document the deployment behavior of the reference implementation.

The benchmark uses batch size one, $3 \times 256 \times 256$ inputs, 25 warmups and 75 CPU / 200 GPU timed repetitions. Table 16 reports means and standard deviations. Speedup is FP32 latency divided by method latency; values below one indicate slower execution. CPU timing variance is substantial, so small differences between routes should not be overinterpreted. The recorded GPU is an NVIDIA H100 80GB HBM3; the CPU model and thread configuration are not recorded. These timings should not be extrapolated to other platforms.

Serialized model states remain 322,499 bytes on CPU and 324,227 bytes on GPU for every method. Peak GPU allocation is 122,115,072 bytes for FP32 and 129,980,416 bytes for each quantized variant. Neither measured serialization nor peak allocation improves. No energy value was recorded. The deployment storage column describes ideal whole-model packed reduction, unlike the selected-weight reduction in routing summaries (for uniform W6/A7 these are 77.592% and 81.250%, respectively).

D.11 PROTES Validation-Subset Robustness

Because PROTES repeatedly evaluates candidate networks using a small validation subset, we tested whether the refinement process is sensitive to the particular images selected for this inner loop. Five distinct 32-patch validation subsets were evaluated while keeping the QUBO initialization, optimizer seed, target operating point, candidate space, and evaluation budget fixed. Each refined assignment was subsequently evaluated on the complete held-out patch split with shared source images.

Across the five subsets, the resulting static-test PSNR was 36.712 ± 0.037 dB, indicating relatively small variation in final restoration quality despite changes in the inner-loop validation sample. However, the exact routing assignments were less stable: pairwise weight agreement ranged approximately from 0.47 to 0.74, activation agreement from 0.68 to 0.92, and joint weight–activation agreement from 0.37 to 0.58. Thus, the experiment supports stability of achieved restoration quality more strongly than stability of the exact discrete architecture. The subsets overlap (pairwise Jaccard indices 0.231–0.391). The requested target is 4.102%, subset-selection seed 4401 and optimizer

seed 4402. After LSQ+, gains over the QUBO reference range from -0.034 to $+0.033$ dB; the small static gains do not consistently survive recovery.

Table 17: Robustness of PROTES to the choice of the 32-patch inner-loop validation subset. Agreement is computed pairwise across the five resulting routing assignments.

Quantity	**Observed result**
Static test PSNR	36.712 ± 0.037 dB
Weight-bit agreement	0.47–0.74
Activation-bit agreement	0.68–0.92
Joint W+A agreement	0.37–0.58

D.12 Simulated Annealing Multi-Seed Robustness

Because the QUBO stage uses stochastic Simulated Annealing, we evaluated the same formulation under five independent solver seeds (0–4) at three representative compute penalties: $\gamma \in \{1.584893, 11.659144, 85.77\}$. Each run used 500 annealing reads and the same profiling data and structural hyperparameters. We report the mean and standard deviation of the resulting held-out-patch PSNR, BOP ratio, and QUBO energy across the five seeds.

The resulting PSNR means were 35.602±1.146 dB, 34.746±1.156 dB, and 34.930±0.389 dB for $\gamma = 1.584893$, 11.659144, and 85.77, respectively. The corresponding BOP ratios were 4.109±0.366%, 4.070±0.543%, and 3.477±0.112%. QUBO energy variation was small within each penalty regime, with standard deviations of 0.043, 0.026, and 0.086, respectively.

However, agreement between the exact discrete routings was substantially lower than the stability of the objective values might suggest. Mean pairwise coordinate agreement over the concatenated weight and activation vectors was 0.304, 0.281, and 0.311 across the three penalty regimes. Weight-only agreement was 0.190, 0.163, and 0.238, while activation-only agreement was 0.418, 0.400, and 0.385. This coordinate agreement is the average of weight and activation agreement; it differs from the both-precisions-at-one-layer agreement used in the subset study. Thus, different annealing seeds can produce materially different discrete routes while yielding close surrogate energies and, in two regimes, PSNR standard deviations exceeding 1 dB. This result motivates reporting stochastic variation rather than treating a single annealing realization as a uniquely identified architecture.

Table 18: Five-seed Simulated Annealing robustness across three representative QUBO compute penalties. Values are mean ± standard deviation over five solver seeds.

γ	Test PSNR (dB)	BOPs (%)	QUBO energy
1.584893	35.602 ± 1.146	4.109 ± 0.366	-321.972 ± 0.043
11.659144	34.746 ± 1.156	4.070 ± 0.543	-331.907 ± 0.026
85.77	34.930 ± 0.389	3.477 ± 0.112	-405.460 ± 0.086

D.13 PROTES Seed Variation

We additionally run three PROTES seeds at each of three all-convolution QUBO seeds. The saved QUBO route is fixed within each γ row; therefore its zero standard deviation is not evidence of QUBO solver robustness. PROTES uses 2,000 proposals per run. It improves the fixed static QUBO score in all nine runs, with the variation shown in Table 19. The runtime was recorded on the separate server of unknown specification and is reported only to indicate the within-study cost.

Table 19: Three-seed PROTES refinement of fixed QUBO routes. Values are mean ± standard deviation over PROTES seeds.

γ	QUBO PSNR	PROTES PSNR	ΔPSNR	PROTES BOPs (%)
1.584893	34.231	36.149 ± 0.057	1.918 ± 0.057	3.962
11.659144	33.513	36.023 ± 0.225	2.509 ± 0.225	3.702
85.77	33.495	35.565 ± 0.289	2.069 ± 0.289	3.409

The mean per-run search times for the three rows are 2,274, 2,365 and 2,272 s, respectively. Exact routes change substantially: the mean number of layers with a different joint weight–activation choice is 24.7–30.0 out of 40.

D.14 Sensitivity to Topological and Objective Hyperparameters

The cross-layer interaction coefficient ω and QUBO objective scaling coefficient β control the relative contribution of structural interactions and task-aware damage terms. We therefore evaluate $\omega \in \{0, 0.05, 0.10, 0.15, 0.20, 0.30, 0.50\}$ and $\beta \in \{0.5, 1, 2\}$ at two representative compression regimes, $\gamma = 11.66$ and $\gamma = 85.77$.

All sweep quality values are evaluated on the test loader. The following maxima are descriptive exploratory observations, not validation-selected hyperparameters or independent evidence for choosing ω.

The response is non-monotonic in both parameters. At $\gamma = 11.66$, PSNR ranges from 33.46 to 35.60 dB across the tested combinations, while BOPs range from 3.66% to 4.59%. At $\gamma = 85.77$, PSNR ranges from 32.89 to 35.76 dB and BOPs from 3.27% to 4.09%.

The best observed configuration in the sweep at $\gamma = 11.66$ is $(\omega, \beta) = (0.05, 1)$, giving 35.604 dB at 4.210% BOPs. At $\gamma = 85.77$, the best observed configuration is $(\omega, \beta) = (0.15, 2)$, giving 35.759 dB at 4.090% BOPs. These results show that the effect of cross-layer interaction depends on the compression regime and objective scaling. We retain $\omega = 0.15$ as the preliminary setting, without treating this test-set sweep as a valid selection procedure or evidence of a universal optimum.

Table 20: Observed sensitivity to the cross-layer coefficient ω and objective scaling coefficient β at two representative compute penalties. The reported configuration is an empirically selected setting, not a claim of a universal optimum.

γ	**PSNR range (dB)**	**BOPs range (%)**	**Best observed** (ω, β)
11.66	33.46–35.60	3.66–4.59	(0.05, 1)
85.77	32.89–35.76	3.27–4.09	(0.15, 2)

PROTES was then run from 30 saved QUBO assignments covering five values of ω, all three β values and both γ regimes. At $\gamma = 11.66$, refinement improves all 15 assignments; at $\gamma = 85.77$, it improves 14 of 15. Table 21 summarizes the full extension. These are single PROTES seeds per configuration and the sweep uses held-out-patch quality descriptively, so it should not be read as a tuned generalization estimate.

Table 21: PROTES extension of the ω–β sweep. Gain is relative to the corresponding saved QUBO route over 15 configurations per row.

γ	PROTES PSNR range	Median gain	Gain range (dB)
11.66	35.832–36.411	+1.518	+0.443 to +2.556
85.77	34.993–36.000	+1.595	-0.350 to +2.311

D.15 Robustness of Empirical Activation Sensitivity to Probe Precision

Empirical Global Activation Sensitivity uses a temporary quantization probe to estimate the downstream effect of perturbing an intermediate activation. To test whether the resulting sensitivity ranking depends strongly on that probe precision, we repeated the profiling procedure using 3-, 4-, 5-, and 6-bit probes.

The pairwise Spearman correlations of the resulting layer rankings range from 0.936 to 0.984, indicating strong agreement in the relative ordering of layer sensitivities. Pearson correlations are lower, ranging from 0.698 for the 3-bit versus 6-bit comparison to 0.967 for the 3-bit versus 4-bit comparison. Thus, rankings are more stable than the absolute sensitivity scale. Because coefficient magnitudes also enter the QUBO, rank agreement alone does not establish identical routing or performance across probe precisions.

Table 22: Pairwise correlations between activation-sensitivity rankings obtained with different probe precisions.

Probe pair	**Spearman**	**Pearson**
3 vs. 4 bits	0.978	0.967
3 vs. 5 bits	0.944	0.844
3 vs. 6 bits	0.936	0.698
4 vs. 5 bits	0.984	0.952
4 vs. 6 bits	0.969	0.852
5 vs. 6 bits	0.983	0.966

D.16 Activation Search-Space Ablation

The main search restricts activation candidates to $\{6, 7, 8\}$. To directly test the motivation for this restriction, we compare it with an otherwise identical search that allows $\{4, 5, 6, 7, 8\}$.

The unrestricted search exhibits a pronounced activation cliff. Its PSNR is 30.12 dB at $\gamma = 1$ and 26.10 dB at $\gamma = 50$. PROTES recovers to 36.04 and 35.15 dB, respectively, while moving the first route from 3.569% to 4.410% BOPs. Thus, the cliff is not intrinsic to the expanded activation menu, but direct refinement can spend additional compute to escape it. At $\gamma = 50$, it improves quality by 9.05 dB while slightly reducing BOPs. The comparison still uses one refinement seed per row and test quality is reported descriptively.

Table 23: Restricted versus unrestricted activation menus before and after PROTES refinement.

Activation space	γ	QUBO BOPs	QUBO PSNR	PROTES BOPs	PROTES PSNR
$\{6, 7, 8\}$	1	4.482	35.126	4.107	36.111
$\{6, 7, 8\}$	5	4.211	35.398	4.085	36.359
$\{6, 7, 8\}$	50	3.237	33.426	3.419	36.146
$\{6, 7, 8\}$	200	3.287	33.772	2.848	34.756
$\{4, 5, 6, 7, 8\}$	1	3.569	30.116	4.410	36.045
$\{4, 5, 6, 7, 8\}$	5	3.405	32.990	3.561	35.484
$\{4, 5, 6, 7, 8\}$	50	3.527	26.101	3.362	35.151
$\{4, 5, 6, 7, 8\}$	200	2.832	29.293	2.665	33.735

D.17 Bias Precision Ablation

The main experiments retain convolutional biases in their original precision. To quantify the impact of this choice, we compare FP32 and INT8 bias quantization for three representative uniform assignments. INT8 bias quantization changes PSNR by -0.145 dB for Uniform W5/A7, -0.036 dB for Uniform W6/A7, and -0.128 dB for W4/A8. The corresponding SSIM changes are small.

For two refined mixed-precision assignments, INT8 bias evaluation instead changes PSNR by +0.0767 and +0.0046 dB. These small sign-changing differences are within the variation expected

from batch-dependent fake quantization and do not establish an advantage for INT8 biases. The experiment is not an integer-only deployment study. We therefore retain FP32 biases in the main experiments and make this assumption explicit.

Table 24: Effect of INT8 bias quantization on representative assignments. Uniform BOPs are exact analytical products of the displayed weight and activation precisions relative to W32/A32.

Assignment	BOPs (%)	ΔPSNR (dB)	ΔSSIM
Uniform W5/A7	3.418	-0.145	-0.0018
Uniform W6/A7	4.102	-0.036	-0.0004
W4/A8	3.125	-0.128	+0.0001
QUBO+PROTES, $\gamma = 11.66$	3.952	+0.0767	+0.0008
QUBO+PROTES, $\gamma = 85.77$	3.140	+0.0046	+0.0003

D.18 Computational Search Cost

The protected-layer formulation contains 304 binary variables and 1,713 non-zero coefficients at the representative target. Its 512 exclusion pairs comprise 494 within-layer one-hot conflicts and 18 skip-mismatch pairs. There are also 570 W–A BOP couplings and 333 ordered-layer interaction pairs; some keys overlap, so these counts are not disjoint. The all-convolution variant has 320 variables and 1,803 coefficients.

Table 25: Offline costs at the 4.102% primary requested target. Shared stages are separated from target-specific search. Cache entries in the original PROTES search count sampled index vectors, not necessarily unique routes.

Stage	Time (s)
Weight-gradient profiling	0.387
Candidate activation profiling	0.233
Global activation profiling	4.834
LSQ+ activation initialization	20.488
QUBO component construction	5.074
QUBO gamma search + independent 500-read solve	1.819
HAWQ-style Hessian profiling	2.568
HAWQ-style knapsack allocation	0.059
Original PROTES refinement	529.075

The original refinement record has 4,890 cached index evaluations at this target, approximately 0.108 s per entry, and a recorded nominal budget of 5,000. This differs from the 2,000-budget repeated-search ablation. The recorded dense-checkpoint-to-final-LSQ+ times are 185.319 s for uniform, 205.565 s for HAWQ-style allocation, 198.140 s for QUBO and 734.474 s for QUBO + PROTES. The corresponding LSQ+ training/materialization stages take 164.831, 177.986, 165.304 and 172.564 s. Totals include costs attributed by the originating runs and exclude dense-model training; they should not be reconstructed by summing every shared stage for every method.

Lower inference BOPs do not imply cheaper allocation. Refinement adds hundreds of seconds here for a modest post-QAT improvement; QUBO alone or Hessian–knapsack allocation remains a practical alternative. These timings describe the recorded environment, not hardware-independent complexity.

D.19 Exploratory Unmatched-Protocol Generalization Pilot

An older single-seed study used a separately trained NAFNetSmall on a source-image-disjoint SIDD partition and MobileNetV2 on CIFAR-10. Its data, architectures, achieved costs and recovery protocol do not match the main comparison, so the results are excluded from pooled claims. For transparency, NAFNetSmall changes from 33.703 dB for recovered QUBO at 4.113% BOPs to 34.079 dB for recovered QUBO+PROTES at 4.107%; MobileNetV2 changes from 91.680% accuracy at

3.493% BOPs to 91.860% at 3.475%. Uniform routes reach higher recorded scores at other costs in both experiments. These single-seed observations therefore do not support a universal cross-task refinement claim.

D.20 Ablation Studies: Negative Results

We additionally evaluated several alternative sensitivity and routing strategies:

- **Second-Order Hessian Trace Estimation:** At $\gamma = 200.0$, a Hessian-aware QUBO using Hutchinson's estimator underperformed the first-order sensitivity formulation in the evaluated setting.
- **LSQ+ Micro-Calibration:** Optimizing the step size using 50 Adam iterations on isolated tensors before MSE evaluation resulted in a PSNR of 33.427 dB in that preliminary configuration. The expanded experiments nevertheless use a calibrated-weight-error variant; this single result does not establish that calibration is generally inferior.
- **Sub-Layer Block-Wise Routing:** Channel-group routing based on sequential slicing, logical variance sorting, and activation-correlation clustering performed poorly in the evaluated experiments and increased optimization complexity.
- **Dynamic Distributional Entropy:** Using activation entropy to unlock aggressive $\{4, 5\}$-bit choices caused substantial degradation in the tested configurations, indicating that activation sparsity alone was not a reliable indicator of quantization robustness.
- **Anti-Correlation Splitting:** Round-robin distribution of correlated channels across logical blocks produced unstable results in the evaluated experiments.

These results should be interpreted as empirical comparisons among the specific alternatives evaluated in this study rather than as evidence that the tested alternatives are universally inferior.