WHEN MORE SEARCH REDUCES COVERAGE: POPULATION-AWARE DIFFUSION SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

More verifier-guided diffusion search need not improve the generated population. On class-conditional ImageNet, increasing branching beyond an initial gain raises Top-K ADM precision while reducing recall and worsening FID. Retained paths become parents of later candidates, linking selection to future exploration. We investigate this feedback with retention-schedule interventions and introduce Moment-QUBO, a rule that balances verifier quality with matching the candidate population's first two Inception-feature moments. Applying it throughout search improves FID beyond replacing only the final selection, showing that early retention decisions matter beyond terminal reranking. Across two independently seeded SiT-XL/2 curves and a DiT-XL/2 transfer, its advantage over Top-K widens with branching at matched denoiser evaluations. At the largest SiT budget, Moment-QUBO reduces FID by over 32% relative to Top-K in both runs. Benefits depend on representation and operating point: CLIP-CMMD favors determinantal selection in the matched-precision SiT comparison, and ordinary DiT sampling offers lower FID and observed cost at lower ADM precision. These findings identify population retention as a design choice for mitigating adverse scaling in repeated diffusion search.

1 INTRODUCTION

A larger search budget gives a diffusion model more candidates to choose from. Yet the selected images need not form a more representative population (Figure 1). Within our repeated-branching SiT search on ImageNet-256, FID initially improves as each parent produces two candidates instead of one. Increasing branching further to four raises Top-K precision but reduces recall and worsens FID from 5.456 to 9.556 (Figure 2). We ask how the choice of retained paths changes the value of additional search compute.

In repeated branching, each retained path generates the next round's candidates. Pointwise ranking can assign many slots to similar high-scoring paths, directing later compute toward their descendants. A pruning decision therefore changes both the current population and the inputs to subsequent search. We ask whether controlling this feedback improves the outcome beyond choosing better final images.

We use candidate-moment matching as a simple intervention. Moment-QUBO balances verifier quality against discrepancies between the retained and candidate populations' first two Inception-feature moments. Each new pool supplies the reference, encouraging the retained parents to represent what search has discovered. At fixed cardinality, this yields a quadratic set objective. Noise search and verifier hacking were studied by Ma et al. (2025); Section 2 distinguishes our population target and repeated branching from existing group-selection methods.

On an independent SiT noise population at the largest budget, replacing only the final selection reduces FID from 9.525 to 7.763; using moment retention throughout search reaches 6.455 (Figure 3). Two SiT curves and a DiT transfer then test how this advantage changes with branching. Comparisons with other population objectives, ordinary sampling and cross-representation metrics establish where the intervention helps and what it trades off.

Our primary contribution is to study retention as an allocation of future search compute. We (i) document adverse compute-scaling trends under pointwise retention; (ii) use round interventions to distinguish final selection from the effects of earlier retention; and (iii) evaluate candidate-moment

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

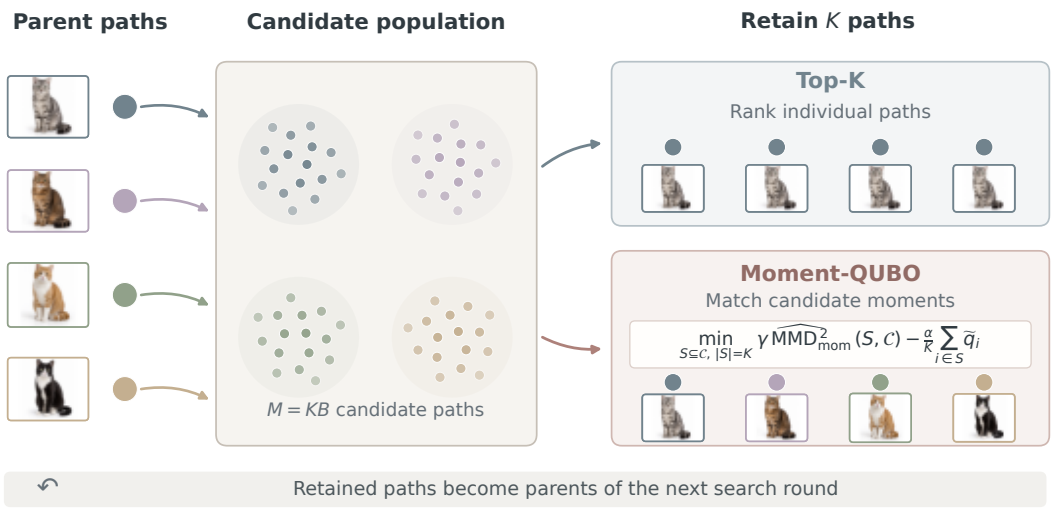

Figure 1: Retention shapes future exploration. Top-K ranks paths independently; Moment-QUBO balances quality and candidate-moment matching. Survivors seed the next round. Cats and colors denote schematic feature groups; K is fixed without per-parent quotas.

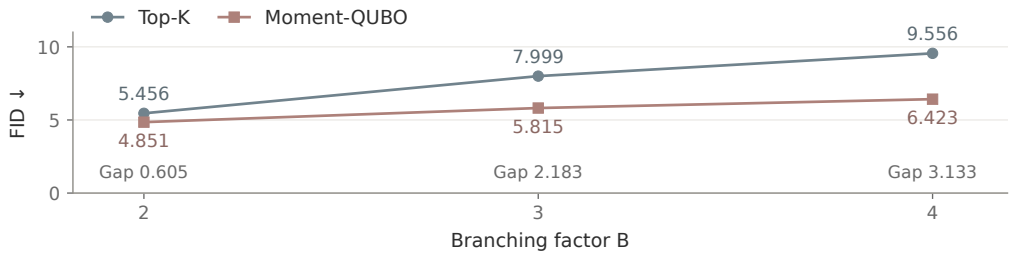

Figure 2: SiT-original (SiT-XL/2, ImageNet-256), with 50K images per point. Gap denotes Top-K minus Moment-QUBO FID, widening from 0.605 at $B = 2$ to 3.133 at $B = 4$. NFE is matched between selectors at each B and increases across B. The FID axis starts at zero. Complete curves appear in Figure 4.

matching across independent noise populations and two backbones, using calibrated baselines to characterize quality, coverage, cost and representation-dependent trade-offs.

2 Why Pointwise Retention Can Concentrate a Population

Pointwise search allocates each retained slot to the highest available verifier score. To see what this implies for a population, let candidate i at a pruning round have latent state z_i, predicted-image features f_i, and verifier score q_i. Retaining the top K candidates solves

$$\max_{x\in\{0,1\}^M} \sum_i q_i x_i \quad \text{s.t.} \quad \mathbf{1}^\top x = K. \tag{1}$$

Each candidate contributes the same score regardless of its companions. Thus, Equation 1 values a high-scoring path equally whether its feature region is new to the retained set or already well represented.

A two-mode example isolates this effect of ranking. Suppose candidates independently fall in a verifier-preferred mode with probability $0 < p < 1$, and every candidate in that mode outranks every alternative. With K fixed, the probability that all selected candidates belong to that mode approaches one as M increases. Appendix A.3 gives the calculation and the condition under which a pairwise

diversity term favors retaining another mode. We test this population-dependent view through lineage measurements and interventions on which rounds use joint selection in Section 4.

Relation to population search. Population control includes particle potentials, MMD gradients and variance-aware resampling (Corso et al., 2024; Mahdizadeh Sani et al., 2026; Shekhar et al., 2026). Group Inference already uses a quadratic quality–diversity objective to prune intermediate candidates (Parmar et al., 2026). Our rule matches the moments of the current candidate pool during repeated branching, where retained paths become parents of subsequent candidates. Appendix B compares the population targets and candidate-evolution rules.

3 Population-Aware Path Selection

Moment-QUBO replaces the retention step in Search over Paths (Ma et al., 2025). Starting from K Gaussian noise samples, we integrate to the first search time, then repeat the following procedure within each class:

One round of population-aware path search

1. Re-noise each of the K parents B times and briefly denoise each branch, giving $M = KB$ candidate latent states z_i.
2. Decode their predicted clean latents into images to obtain scores q_i and population features f_i. Set the reference to $Y = \{f_i\}_{i=1}^{M}$.
3. Jointly select K candidates by balancing verifier quality and moment discrepancy from Y using the objective below.
4. Retain the selected latent states as the next parents. At the clean endpoint, return the selected images.

The first three rounds retain intermediate states that become the next parents; the fourth selects final images. Thus each early decision affects which candidates will be available later. Appendix B specifies the times, re-noising transition and clean prediction.

3.1 A population target for each pruning decision

Each expansion produces a population of possible continuations. We use this pool as a local reference: matching its statistics limits the additional distortion introduced by pruning, while the quality term favors high-scoring candidates. Refreshing the reference after each expansion lets the target follow discoveries made during search. Let $Y = \{y_r\}_{r=1}^{R} = \{f_i\}_{i=1}^{M}$ denote these features, with $R = M$, and let k be a positive semidefinite kernel. The biased empirical estimate of squared maximum mean discrepancy (MMD) for a size-K selection is (Gretton et al., 2012)

$$\widehat{\mathrm{MMD}}^2(x, Y) = \frac{1}{K^2}\sum_{i,j} k(f_i, f_j)x_i x_j - \frac{2}{KR}\sum_i \sum_r k(f_i, y_r)x_i + C_Y. \tag{2}$$

Here $C_Y = R^{-2}\sum_{r,s} k(y_r, y_s)$ is independent of x. We minimize squared MMD while rewarding robustly standardized verifier quality, using $\alpha, \gamma \geq 0$ and binary x with $\mathbf{1}^\top x = K$:

$$E_{\mathrm{MMD}}(x) = \gamma\widehat{\mathrm{MMD}}^2(x, Y) - \frac{\alpha}{K}\sum_i \widetilde{q}_i x_i. \tag{3}$$

This objective regularizes each retention decision relative to its input population. Bias already present in the candidates carries into the reference; rediscovering an absent mode depends on subsequent branching and denoising.

The kernel specifies which population statistics matter. FID compares Gaussian approximations to Inception feature distributions (Heusel et al., 2017). Its covariance square-root term is nonquadratic in the selection variables. To obtain a quadratic first/second-moment proxy, we use the reference-centered kernel

$$k_{\mathrm{mom}}(u, v) = w_1\frac{(u-c)^\top(v-c)}{s_1} + w_2\frac{\big((u-c)^\top(v-c)\big)^2}{s_2}, \tag{4}$$

where c is the reference mean, $s_1, s_2 > 0$ normalize the terms, and $w_1, w_2 \geq 0$ have positive sum. Moment-QUBO instantiates the MMD-QUBO framework with this moment kernel. The moment discrepancy matches means and second raw moments centered by c and is available before denoising completes. Dynamic matching targets the current candidates; final FID targets real data. The objective is an FID-motivated proxy rather than exact FID optimization. Appendix A gives the expansion; Table 4 distinguishes the references and kernels used by the RBF controls and global-selection extensions.

3.2 A PATH'S VALUE DEPENDS ON THE RETAINED SET

For fixed K, the MMD denominators are constants and $x_i^2 = x_i$. Equation 3 therefore has the form

$$E(x) = \sum_i a_i x_i + \sum_{i<j} b_{ij} x_i x_j, \qquad \mathbf{1}^\top x = K. \tag{5}$$

The linear coefficients combine quality and attraction to the reference; the pairwise coefficients couple candidates through their feature statistics (Equation 11). Nonnegative kernel entries penalize redundant pairs; centered moment-kernel entries can have either sign.

For a retained set S, replacing $r \in S$ with $u \notin S$ changes the objective by

$$\Delta_{r \to u} E(S) = a_u - a_r + \sum_{j \in S \setminus \{r\}} (B_{uj} - B_{rj}), \tag{6}$$

where $B_{ij} = b_{\min(i,j),\max(i,j)}$ for $i \neq j$. A candidate's exchange value depends on the other retained paths; for $K = 1$, this interaction disappears. In dynamic search, we optimize directly over size-K selections. For unconstrained solver backends, adding $A(\mathbf{1}^\top x - K)^2$ with sufficiently large A yields a quadratic unconstrained binary optimization (QUBO) formulation (Glover et al., 2022).

Optimization. The dynamic solver constructs a size-K set greedily using marginal objective costs, then applies the best energy-decreasing one-out/one-in exchange per scan, always preserving cardinality. Refinement stops at tolerance 10^{-12} or 100 scans. The largest dynamic problems have $M = 200$ and $K = 50$; Appendix A.4 gives the algorithm and complexity.

4 DOES RETENTION CHANGE FUTURE EXPLORATION?

Common setting. We use SiT-XL/2 (Ma et al., 2024) on ImageNet-256 (Deng et al., 2009; Russakovsky et al., 2015), with $K = 50$ paths per class, four retention rounds and $B \in \{1, 2, 3, 4\}$. Within each curve, selectors share initial states and branch-noise indices at matched denoiser function evaluations (NFE); their later candidates diverge when they retain different parents. The two curves, *SiT-original* and *SiT-repeat*, use independent noise seeds. Online quality is the target-class logit from a DINOv2 ViT-L/14 linear head (Oquab et al., 2024). Moment-QUBO uses 2048-dimensional Inception-v3 features (Szegedy et al., 2016), $w_1 = w_2 = \gamma = 1$ and $\alpha = 0.56$ throughout search. A 16-class development screen chose the quality weight before formal ADM evaluation; the original DINO-based determinantal point process (DINO-DPP) (Kulesza & Taskar, 2012) baseline used fixed $\eta = 2$. Appendix B details this unequal tuning effort, noise sharing and sampling schedules.

Evaluation. Each full evaluation contains 50 images from each of 1,000 classes (50K images). The ADM pipeline (Dhariwal & Nichol, 2021) reports FID, spatial FID (sFID), Inception Score (IS) (Salimans et al., 2016) and precision/recall (Kynkäänniemi et al., 2019). ADM precision measures membership in the estimated real-data feature manifold after generation; it differs from the online verifier. ADM recall, pairwise feature distance and lineage entropy describe feature-space coverage, sample separation and retained ancestry. DINO metrics and CLIP-CMMD (Jayasumana et al., 2024) test sensitivity to representation. Paired conditional 95% confidence intervals (CIs) measure class/query composition sensitivity for fixed outputs and reference; independent generations separately test noise replication.

4.1 INTERVENING ON THE RETENTION ROUNDS

We first ask whether early retention affects the final population beyond terminal selection. We change which rounds use joint selection. On SiT-repeat at $B = 4$, switching from Top-K to the main

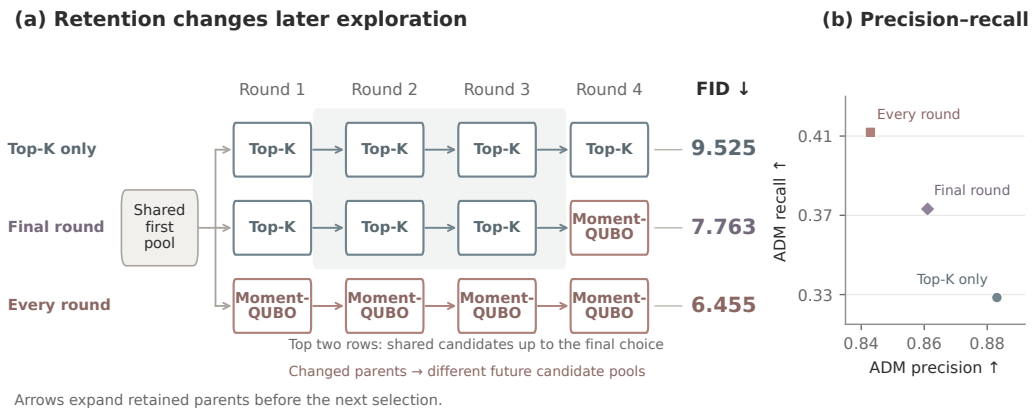

Figure 3: Retention-schedule intervention on SiT-repeat at $B = 4$, with 50K images per policy. (a) Top-K only and final-round Moment-QUBO share candidate pools through the final pruning decision. Every-round Moment-QUBO changes the parents from round 1, so later pools can differ. The policies share the denoising budget; final-round replacement reuses fully decoded candidates and adds zero diffusion NFE. (b) ADM precision–recall shows the corresponding operating points.

Moment-QUBO rule only at the last round lowers FID from 9.525 to 7.763; using it in all four rounds reaches 6.455 (Figure 3; Table 11). Both QUBO policies use the same $\alpha = 0.56$ objective and solver. The last-round intervention reuses fully decoded Top-K candidates and adds zero diffusion NFE. Its improvement isolates a final-selection effect; the complete-policy result additionally reflects changed parents and subsequent candidate populations.

Parent identities provide a complementary view of this process. Across 1,000 classes in the principal $K = 50$ pipelines, final-round $B = 4$ lineage entropy is higher under Moment-QUBO than Top-K by 0.0309/0.0339 on SiT-original/SiT-repeat, and pairwise DINO distance increases by 2.260/2.226. Conditional 95% class-bootstrap CIs exclude zero for all four differences (Appendix C.2). Entropy declines over rounds for all selectors, with joint pruning preserving more ancestry. The separate $K = 4$ trees and images illustrate this process (Appendix G).

The retention effect also appears with CLIP cosine similarity as the online verifier. On 128 prospectively selected classes at $B = 4$, QUBO raises DINO distance by 2.708 and root entropy by 0.0328 over Top-K, while CLIP score falls by 0.69%. All three conditional 95% CIs exclude zero; each configuration returns 6,400 images (Appendix C.17).

5 Does the Effect Grow with Search Compute?

Having established that early retention matters, we compare complete retention policies as branching increases. In SiT, $B = 1$ still runs all four re-noising rounds and retains every candidate; ordinary sampling is a separate baseline. The denoiser evaluations per output are $222 + 43B \in \{265, 308, 351, 394\}$. Budgets share initial states and nested branch-noise indices within each curve. Both SiT curves extend previously known $B = 4$ endpoints (Appendix B).

5.1 Independent SiT noise populations

SiT-original initially benefits from additional branching within the re-noising procedure: all selectors coincide at $B = 1$ (FID 16.590), and FID improves at $B = 2$. Further branching separates precision from coverage. From $B = 2$ to $B = 4$, Top-K precision rises from 0.8352 to 0.8829, but recall falls from 0.4574 to 0.3302 and FID increases from 5.456 to 9.556 (Figure 4, Table 13). The adverse trend thus follows an initial search gain.

Changing the retention objective reduces this deterioration. Relative to Top-K, Moment-QUBO yields lower FID and higher recall at every $B > 1$ on this SiT curve; its FID advantage grows from 0.605 to 2.183 to 3.133 at $B = 2, 3, 4$. At $B = 4$, it obtains 6.423 FID and 0.4026 recall at the same

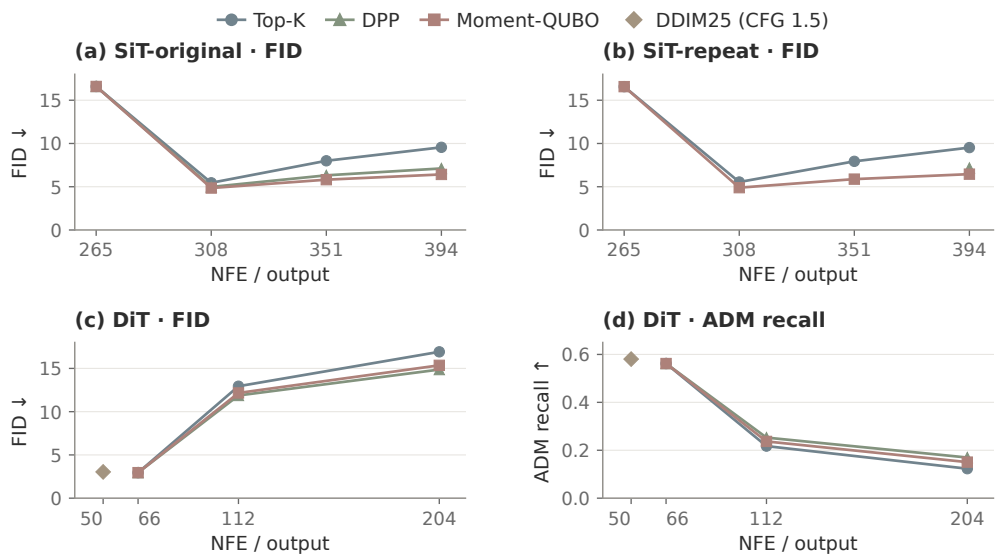

Figure 4: Repeated branching on two backbones, with 50K images per point. (a,b) Independently seeded SiT-original and SiT-repeat. $B = 1, 2, 3, 4$ correspond to 265, 308, 351, 394 NFE/output. Within each curve, selectors share the $B = 1$ population; SiT-repeat separately evaluates DPP only at $B = 4$, so no intermediate DPP points are inferred. (c,d) DiT FID and recall: $B = 1, 2, 4$ correspond to 66, 112, 204 NFE/output, with re-noising included at $B = 1$. Diamond markers show ordinary DDIM25 (CFG 1.5) at 50 NFE/output. DPP uses DINO features on SiT and Inception features with $\eta = 4$ on DiT. Within each study, budgets share initial states and nested branch-noise indices.

394 denoising NFE/output. Its own FID still rises from 4.851 to 6.423 and recall falls from 0.4914 to 0.4026 between $B = 2$ and $B = 4$. The benefit is therefore mitigation under larger branching. The $B = 4$ recall gain of 0.0724 comes with a precision reduction of 0.0355, exceeding the pre-specified 0.02 precision-loss threshold.

SiT-repeat reproduces the widening FID advantage: 0.653, 2.056, and 3.070 at $B = 2, 3, 4$, with recall gains 0.0397, 0.0606, and 0.0834. At $B = 4$, QUBO reaches 6.455 FID versus 9.525 for Top-K and 7.172 for DINO-DPP (Table 15).

On SiT-original at $B = 4$, conditional 95% CIs exclude zero for QUBO's FID and recall gains and precision loss versus Top-K. Against DINO-DPP, only the FID CI excludes zero (Appendix C.10).

5.2 Transfer to a second backbone

We transfer the objective to four-round DiT-XL/2 (Peebles & Xie, 2023) search with $B \in \{1, 2, 4\}$ and $2(10 + 23B)$ NFE/output. After a 32-class pilot, eight frozen configurations each generate 50K new images (Appendix E.2). This experiment reproduces the trend (Figure 4c,d). At $B = 2, 4$, Moment-QUBO obtains FID 12.162/15.356 versus 12.941/16.916 for Top-K: the gap grows from 0.779 to 1.560, with recall gains of 0.0195/0.0274. All four conditional 95% CIs exclude zero. Inception-DPP with $\eta = 4$ attains lower FID (11.873/14.866) at lower precision; Section 6 separately calibrates its quality weight. These results support mitigation relative to Top-K on a second backbone. Ordinary DDIM25 at CFG 1.5 achieves 3.036 FID and 0.5806 recall, outperforming all tested $B \geq 2$ search configurations on both metrics (Appendix E.2). Our retention study therefore concerns the higher-ADM-precision regime induced by verifier-guided search.

6 Comparing Population Objectives

The interventions and budget curves motivate joint retention. We next ask which population target to use: we compare moment matching with DPP at calibrated precision, then use quadratic repulsion

Table 1: SiT-confirm: independent dynamic $B = 4$ comparison with frozen $\alpha = 0.56$ and $\eta = 4$. Both methods generate 50K new outputs, using Inception population features, the same DINO verifier and paired noise indices.

Selector	FID↓	Prec.↑	Rec.↑	CLIP-CMMD↓
Moment-QUBO	6.429	0.845	0.413	0.348
Inception-DPP ($\eta = 4$)	7.516	0.845	0.392	0.335

Table 2: DiT precision calibration and ordinary sampling on a new noise cohort, each with 50K outputs. Search uses $B = 2$, CFG 1.5 and 112 NFE/output; ordinary DDIM25 uses CFG 3 and 50 NFE/output. Time is the observed median warm inference latency per 50 images on shared devices.

Selector	FID↓	Prec.↑	Rec.↑	Time (s)
Moment-QUBO ($B = 2$)	12.162	0.935	0.238	64.78
Inception-DPP ($\eta = 16$)	12.941	0.935	0.223	64.80
DDIM25 (CFG = 3)	9.544	0.912	0.369	23.96

and moment ablations to identify what the statistical target contributes. Solver refinement is evaluated separately.

6.1 Comparing quality–coverage operating points

On SiT-original, Inception-DPP has lower FID and higher recall than Moment-QUBO at both tested budgets (Appendix Table 22). At $B = 4$, DPP reaches 5.451 FID and 0.492 recall with 0.777 precision; QUBO reaches 6.423 FID and 0.403 recall with 0.847 precision. With features matched, determinant repulsion and moment matching occupy different quality–coverage points. We therefore calibrate DPP's quality weight to compare FID and recall at closely matched ADM precision.

With QUBO fixed at $\alpha = 0.56$, screening DPP $\eta \in \{4, 5, 6\}$ selects $\eta = 4$. The estimated absolute precision difference is at most 0.005 on both the 128-class screening subset and the separate held-out subset. After a same-noise extension (Appendix C.14), SiT-confirm tests the two frozen policies on 50K new images per method, using new initial and branch-noise seeds and all registered metrics (Appendix C.16).

On SiT-confirm, QUBO–DPP precision differs by $+0.00016$ (conditional 95% CI $[-0.00440, 0.00494]$, within ± 0.005). QUBO lowers FID by 1.087 (14.5%) and raises recall by 0.0203; paired conditional 95% CIs are $[-1.522, -0.713]$ and $[0.0105, 0.0298]$. CLIP-CMMD favors DPP by 0.01256 (conditional 95% CI $[0.00910, 0.01584]$). Evaluation on new noise seeds reproduces the Inception FID/coverage benefit and the cross-representation trade-off. Broader DINO/CLIP audits show representation-dependent orderings, with Top-K favored by several cross-representation metrics (Appendix F).

For DiT, a 32-class screen selects DPP $\eta = 16$ at $B = 2$, and 64 disjoint classes check the frozen choice. The qualifying pair is then tested on 50K new outputs per method, yielding the same Inception-space ordering (Table 2). With DPP fixed at $\eta = 16$, the precision difference is $+0.00004$ (conditional 95% CI $[-0.00340, 0.00352]$, inside ± 0.005). Moment-QUBO lowers FID by 0.778 (6.0%; conditional 95% CI $[-1.160, -0.420]$) and raises recall by 0.0146 ($[0.0060, 0.0232]$). CLIP-CMMD is 0.19450 versus 0.19310; its conditional 95% difference CI $[-0.00054, 0.00328]$ spans zero. Ordinary CFG 3 sampling reaches lower FID and higher recall with less observed time, at precision lower by 0.02252. Appendix J reports all calibration gates, the subsequent ordinary-sampling evaluation and conditional 95% CIs. The SiT and DiT comparisons establish Inception-space gains at the tested ADM-precision points. In DiT, the value of higher ADM precision over ordinary sampling depends on whether it warrants added cost and lower ADM recall. The frozen-pool sweep in Appendix C.12 separately examines selection from fixed completed candidates.

Table 3: Independent SiT retention confirmation at $B = 4$: 50K new images per frozen policy. Initial and branch noise are independent of development and the earlier same-noise extension. Full metrics and conditional 95% CIs appear in Appendix E.1.

Selector	FID↓	Prec.↑	Rec.↑
Quadratic diversity (GI-style)	7.496	0.846	0.398
Mean-only matching	6.653	0.846	0.395
Moment-QUBO	6.568	0.846	0.408

6.2 Which population statistics matter?

To separate moment matching from generic pairwise diversity, a Group-Inference-style baseline combines verifier quality with quadratic cosine-distance repulsion inside the same four-round SiT search, with matched Inception features and solver. We calibrate its weight on 16 classes to match QUBO's first-round retained verifier score. Following pilot and full-class evaluations, an independent run generates 50K new images per frozen policy. This tests the objective within our framework; original Group Inference also differs in verifier, candidate evolution and solver (Parmar et al., 2026).

On independent noise, Moment-QUBO lowers FID by 0.928 relative to quadratic repulsion (conditional 95% CI $[-1.217, -0.656]$) and raises recall by 0.0106 ($[0.0014, 0.0196]$); the precision gap is -0.00016. Matching only the mean captures most of the FID gain: adding second moments changes FID by -0.085 ($[-0.255, 0.121]$) and recall by $+0.0131$ ($[0.0048, 0.0213]$; Table 46). The conditional 95% CIs span zero for FID and lie above zero for recall: second moments support a recall benefit, with their incremental FID effect unresolved. Mean matching is itself quadratic in the binary selection variables: the squared mean discrepancy contains pairwise cross terms. It therefore differs from the linear-only ablation. The 64-class MMR, k-medoids and DINO-moment controls further expose representation dependence: DINO-KID favors QUBO over Top-K, whereas CLIP-CMMD favors Top-K (Appendices D.3 and D.4).

Fixed-pool controls further separate statistical matching from ancestry preservation and test the identity of quadratic interactions. Their effects depend on the reference population (Appendices C.4 and C.20).

6.3 How much does refinement contribute?

Greedy selection captures most of the benefit: one-swap improves SiT FID by 0.004/0.042 at $B = 2, 4$. Refining DPP changes its 64-class FID by -0.0011 (Appendix D). Objective design matters more here than deeper optimization.

7 Cost and Scope

7.1 Where computation is spent

The metric comparisons locate the selectors' operating points; complete noise-to-image timing establishes their computational cost. Denoising dominates the SiT pipeline. At $B = 4$, median warm noise-to-image times are 192.06/194.39/193.06 s per 50 outputs for Top-K, Inception-DPP ($\eta = 2$), and QUBO across three shared-device blocks (Figure 5). QUBO construction and optimization total 0.045 s across all four rounds. DiT takes 64.78/64.80 s for QUBO/DPP and 23.96 s for ordinary CFG 3 (Table 2). These times include features, selection and decoding. Shared-device load contributes to variation, so the timings describe observed execution costs under the recorded conditions. Appendices H and J detail component costs and timing scope.

Scope of population control. Candidate matching regularizes pruning relative to the current pool, whose existing bias also enters the reference. The shared Inception representation aligns the objective with ADM metrics; DINO/CLIP audits expose different orderings. Population representation is therefore part of the retention design, alongside the quality criterion and compute budget.

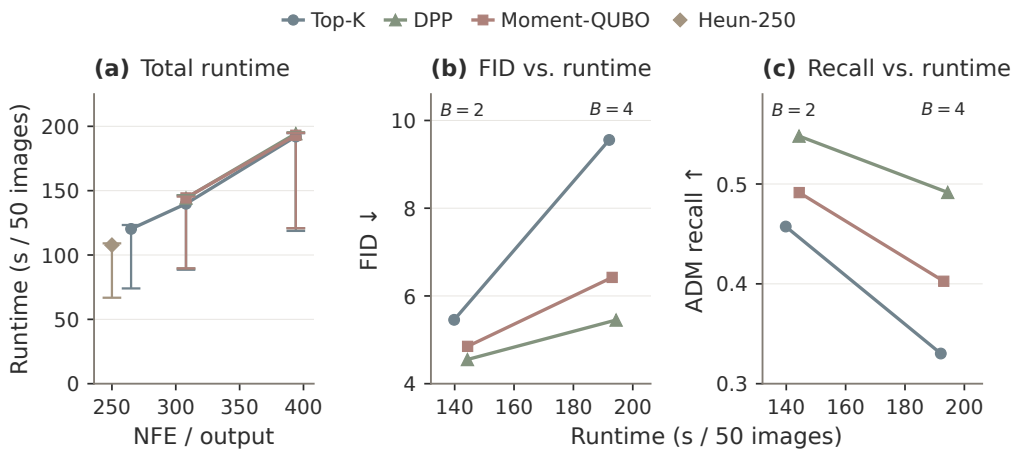

Figure 5: Complete warm inference cost on SiT. (a) Median noise-to-image latency per 50 outputs; capped bars show min–max across three class/device blocks, not confidence intervals. (b,c) Separately measured latency is paired with SiT-original 50K FID/recall at $B = 2, 4$ (left/right). DPP uses Inception features and $\eta = 2$. The Heun-250 diamond has only a timing measurement. Costs include features, selection and decoding; loading, warmup and artifact writes are excluded.

Appendix I additionally demonstrates execution of larger global-selection variants on a physical coherent Ising machine (CIM) (Wei et al., 2026).

8 Conclusion

Retention determines where future search compute is spent. Across two independent SiT curves and a DiT transfer, the FID advantage of candidate-moment matching over Top-K grows with branching. Round interventions show that early retention matters beyond final selection. Independent SiT controls favor moment matching over quadratic repulsion: mean matching captures most of the FID gain, while second moments improve ADM recall.

At the tested ADM-precision points, calibrated DPP comparisons favor moment matching in Inception FID and recall. Ordinary DiT sampling offers lower FID and observed cost at lower precision, while CLIP-CMMD favors DPP on SiT. These ImageNet results make retention an explicit inference-time design choice whose value depends on the quality criterion, population representation and compute budget.

AI Use Statement

AI supported writing and generated the cats in Figure 1. The authors are responsible for all submitted content.

Reproducibility Statement

The appendix specifies objective derivations, frozen protocols and controlled ablations. The source package includes generated tables, compact evidence and source hashes.

References

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1lUOzWCW`.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025. URL https://doi.org/10.1145/290941.291025.

Gabriele Corso, Yilun Xu, Valentin de Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-I.I.D. diverse sampling with diffusion models. In *International Conference on Learning Representations*, pp. 22480–22507, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/612a7948f3294a02a63d970566ca8536-Abstract-Conference.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Fred Glover, Gary Kochenberger, Rick Hennig, and Yu Du. Quantum bridge analytics I: A tutorial on formulating and using QUBO models. *Annals of Operations Research*, 314(1): 141–183, 2022. doi: 10.1007/s10479-022-04634-2. URL https://doi.org/10.1007/s10479-022-04634-2.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Rogério Guimarães and Pietro Perona. Inference-time scaling of diffusion models via progressive seed pruning. In *European Conference on Computer Vision*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/3518.

Haoran He, Jiajun Liang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Ling Pan. Scaling image and video generation via test-time evolutionary search. *arXiv preprint arXiv:2505.17618*, 2025. URL https://arxiv.org/abs/2505.17618.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Samuel Howard and Nikolas Nüsken. A mean-field framework for inference-time distributional control of diffusion models. *arXiv preprint arXiv:2608.08770*, 2026. URL https://arxiv.org/abs/2608.08770.

Takahiro Inagaki, Yoshitaka Haribara, Koji Igarashi, Tomohiro Sonobe, Shuhei Tamate, Toshimori Honjo, Alireza Marandi, Peter L. McMahon, Takeshi Umeki, Koji Enbutsu, et al. A coherent ising machine for 2000-node optimization problems. *Science*, 354(6312):603–606, 2016. doi: 10.1126/science.aah4243. URL https://doi.org/10.1126/science.aah4243.

Vineet Jain, Kusha Sareen, Mohammad Pedramfar, and Siamak Ravanbakhsh. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 145576–145615, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d6484394c4cb5e1f4ecad8d90b912025-Abstract-Conference.html.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/

Jayasumana_Rethinking_FID_Towards_a_Better_Evaluation_Metric_for_Image_Generation_CVPR_2024_paper.html.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990. doi: 10.1002/9780470316801. URL https://doi.org/10.1002/9780470316801.

Scott Kirkpatrick, C. Daniel Gelatt, Jr., and Mario P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983. doi: 10.1126/science.220.4598.671. URL https://doi.org/10.1126/science.220.4598.671.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012. doi: 10.1561/2200000044. URL https://doi.org/10.1561/2200000044.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

Jonathan Lys, Vincent Gripon, Axel Marmoret, Lukas Mauch, Fabien Cardinaux, Ghouthi Boukli Hacene, and Bastien Pasdeloup. D5P4: Partition determinantal point process for diversity in parallel discrete diffusion decoding. *arXiv preprint arXiv:2603.19146*, 2026. URL https://arxiv.org/abs/2603.19146.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40, 2024. doi: 10.1007/978-3-031-72980-5_2. URL https://doi.org/10.1007/978-3-031-72980-5_2.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2523–2534, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_Scaling_Inference_Time_Compute_for_Diffusion_Models_CVPR_2025_paper.html.

Matina Mahdizadeh Sani, Nima Jamali, Mohammad Jalali, and Farzan Farnia. MMD guidance: Training-free distribution adaptation for diffusion models via maximum mean discrepancy guidance. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63518.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search. In *Advances in Neural Information Processing Systems*, volume 38, pp. 13170–13216, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/13b501c58ae3bfe9635a259f4414e943-Abstract-Conference.html.

Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. In *International Conference on Learning Representations*, pp. 89763–89790, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9132fb7713ccc4860399d41695be8622-Abstract-Conference.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

QBOSON. Kaiwu SDK: Classical optimizers, 2026. URL https://kaiwu-sdk-docs.qboson.com/en/source/kaiwu.classical.html. Accessed September 14, 2026; experiments use audited SDK version 1.0.33a.

Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024. URL https://arxiv.org/abs/2407.14041.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Shivanshu Shekhar, Sagnik Mukherjee, Jia Yi Zhang, and Tong Zhang. VASR: Variance-aware systematic resampling for reward-guided diffusion. *arXiv preprint arXiv:2604.06779*, 2026. URL https://arxiv.org/abs/2604.06779.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55810–55827, 2025. URL https://proceedings.mlr.press/v267/singhal25b.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

Harold Szu and Ralph Hartley. Fast simulated annealing. *Physics Letters A*, 122(3–4):157–162, 1987. doi: 10.1016/0375-9601(87)90796-1. URL https://doi.org/10.1016/0375-9601(87)90796-1.

Egor S. Tiunov, Alexander E. Ulanov, and A. I. Lvovsky. Annealing by simulating the coherent ising machine. *Optics Express*, 27(7):10288–10295, 2019. doi: 10.1364/OE.27.010288. URL https://doi.org/10.1364/OE.27.010288.

Hai Wei, Chengjun Ai, Putuo Guo, Bingjie Jia, Lixin Yuan, Hanquan Song, Shaobo Chen, Chongyu Cao, Jie Wu, Chao Ju, Yin Ma, Jintao Fan, Minglie Hu, Chuan Wang, and Kai Wen. A versatile coherent Ising computing platform. *Light: Science & Applications*, 15(1):74, 2026. doi: 10.1038/s41377-025-02178-1. URL https://www.nature.com/articles/s41377-025-02178-1.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Y Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical

search. In *International Conference on Learning Representations*, pp. 39133–39166, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/41c9aa6c04924c29c6469572c5f6d01b-Abstract-Conference.html.

A Derivations and Implementation Details

Guide to the supporting evidence. The main dynamic evidence comprises the SiT budget curves (Tables 13 and 15), independent retention confirmations (Appendix E), and calibrated DPP comparisons (Appendices C.16 and J). Round interventions, ancestry diagnostics and moment ablations explain the retention mechanism (Appendices C.3, C.2 and D.3). Cross-representation and runtime analyses characterize its operating conditions (Appendices F and H). Static selection and physical execution are extensions (Appendices C.8 and I). Calibration histories are documented alongside the resulting frozen-policy evaluations.

Scope of the dynamic comparisons. Candidate matching is a local regularizer on each pruning decision: it balances verifier quality against the change in feature statistics between candidates and survivors. Because the reference is refreshed, the target also reflects biases of the candidates it receives; recovering absent modes depends on subsequent exploration. The calibrated DPP comparisons evaluate frozen policies on new generation noise at closely matched ADM precision. They establish the ordering at those operating points; the shape and extent of the dynamic precision–FID–recall frontier remain a separate question from the fixed-final-pool sweeps. Within each backbone, selectors share a prescribed re-noising schedule and integration budget per branch. The comparisons therefore isolate retention under these settings. Re-noising strength and integration accuracy can also affect candidate quality and parent dependence, so variation across these settings would test how broadly the observed trends transfer.

Method and solver names. MMD-QUBO names the fixed-cardinality MMD-based quadratic formulation. Moment-QUBO uses its first/second-moment kernel; RBF-MMD-QUBO uses an RBF kernel. The main dynamic Moment-QUBO matches the current class candidates. Global moment variants instead match real-reference statistics; the static DiT variant also combines feature blocks and a dispersion term. A move-QUBO reparameterizes a global objective by binary exchanges around a fixed selection, rather than by candidate-inclusion variables. Table 4 locates these variants. Within an explicitly identified study, "QUBO" abbreviates that study's objective.

Moment-greedy and Moment-one-swap optimize the same moment objective; the latter is the greedy-plus-exchange solver used by dynamic Moment-QUBO. RBF-MMD-greedy and RBF-MMD-one-swap analogously denote solvers for the RBF control. Mean-only matching retains the first statistical moment; linear-only deletes binary off-diagonal interactions, so these are different ablations. SA denotes ordinary simulated annealing, FSA denotes the historical fixed-K fast-annealing control, and SimCIM is a numerical simulator; physical CIM denotes the device returns. These solver names specify how an objective is optimized.

Feature and quality labels. DINO-DPP and Inception-DPP identify the features used in the DPP diversity kernel. In the main experiments both still use the DINO class verifier. A common verifier-score floor does not imply equal realized verifier scores or equal ADM precision. "Matched precision" refers to the reported ADM calibration, whereas "quality-floor" identifies a common score constraint. ADM recall, pairwise feature distance and lineage retention describe, respectively, feature-manifold coverage, sample separation and ancestry.

A.1 Expansion of the fixed-cardinality MMD objective

For a selected empirical distribution $\widehat{P}_x = K^{-1}\sum_i x_i\delta_{f_i}$ and reference distribution $\widehat{P}_Y = R^{-1}\sum_r \delta_{y_r}$, the biased squared MMD is

$$\widehat{\mathrm{MMD}}^2(\widehat{P}_x,\widehat{P}_Y) = \frac{1}{K^2}\sum_{i,j} x_i x_j k(f_i,f_j) - \frac{2}{KR}\sum_{i,r} x_i k(f_i,y_r) \tag{7}$$

$$+ \frac{1}{R^2}\sum_{r,s} k(y_r,y_s). \tag{8}$$

Because $x_i^2 = x_i$, split the first sum into diagonal and off-diagonal parts:

$$\frac{1}{K^2}\sum_{i,j} x_i x_j k(f_i,f_j) = \sum_i \frac{k(f_i,f_i)}{K^2}x_i + \sum_{i<j}\frac{2k(f_i,f_j)}{K^2}x_i x_j. \tag{9}$$

Table 4: Roles of the selection formulations and their population targets. Dynamic SiT is the main method.

Study	Population target and formulation	Question tested
Dynamic SiT	Current class candidates; Inception moments	Multi-round retention
Dynamic DiT	Current class candidates; Inception moments	Second branching backbone
Static DiT curves	Real-reference Inception/DINO moments and DINO dispersion	Static budget scaling
DiT RBF control	Current class candidates; DINO RBF	Intermediate-path selection
Final-pruning controls	Current class candidates; Inception moments with a score floor	Last-round retention
Final-pool Pareto	Real-reference global Inception moments	Fixed-candidate trade-off
Frozen-pool CIM	Real-reference global moments; feasible move-QUBO	Physical execution

The reference-only term is constant, yielding the coefficients in Equation 11:

$$a_i = \gamma \left[\frac{k(f_i, f_i)}{K^2} - \frac{2}{KR} \sum_r k(f_i, y_r) \right] - \frac{\alpha}{K} \widetilde{q}_i, \tag{10}$$

$$b_{ij} = \frac{2\gamma}{K^2} k(f_i, f_j), \qquad i < j. \tag{11}$$

The main SiT Moment-QUBO fixes $\gamma = 1$ in every round; its quality weight is $\alpha = 0.56$ and its moment-kernel weights are $w_1 = w_2 = 1$.

We use robust z-scoring

$$\widetilde{q}_i = \frac{q_i - \mathrm{median}(q)}{1.4826\,\mathrm{median}_j\,|q_j - \mathrm{median}(q)|}, \tag{12}$$

with a standard-deviation fallback when the denominator is below 10^{-12}. If that standard deviation is also below 10^{-12}, all standardized scores are set to zero.

A.2 Why the moment kernel matches first and second moments

The linear kernel $k_1(u, v) = u^\top v$ has squared MMD $\|\mathbb{E}_P[u] - \mathbb{E}_Q[u]\|_2^2$. The homogeneous quadratic kernel $k_2(u, v) = (u^\top v)^2$ has feature map $u \otimes u$ and therefore matches $\|\mathbb{E}_P[uu^\top] - \mathbb{E}_Q[uu^\top]\|_F^2$. Centering both distributions by the fixed reference mean c and adding the two kernels gives Equation 4. We set

$$s_1 = \mathbb{E}_{y \sim Y} \|y - c\|_2^2, \tag{13}$$

$$s_2 = \mathbb{E}_{y \sim Y} \|y - c\|_2^4, \tag{14}$$

with a numerical floor of 10^{-12}. The normalization keeps the two terms comparable; both weights are one in the principal SiT experiments.

This construction compares the first and second moments of the retained and reference populations using a quadratic objective that can be evaluated before denoising finishes. FID also compares feature statistics, using squared Euclidean distance for means and the squared Bures distance for covariances: $\mathrm{Tr}[\Sigma_P + \Sigma_Q - 2(\Sigma_P^{1/2} \Sigma_Q \Sigma_P^{1/2})^{1/2}]$.

A.3 A two-mode illustration of selection concentration

Consider a minimal two-mode model in which each of the M candidates belongs independently to a verifier-preferred mode A with probability $0 < p < 1$ or to a second mode B. Suppose every candidate in A has score $q_A > q_B$, and let $N_A \sim \mathrm{Binomial}(M, p)$, with K fixed. Pointwise Top-K selects only mode A whenever $N_A \geq K$, hence

$$\Pr(S_{\mathrm{TopK}} \subseteq A) = 1 - \sum_{j=0}^{K-1} \binom{M}{j} p^j (1-p)^{M-j} \longrightarrow 1 \quad \text{as } M \longrightarrow \infty. \tag{15}$$

Thus, increasing the candidate count can concentrate retention on one mode even with a fixed verifier and exact ranking. The exchange margin of a quadratic set score $F(S) = \alpha \sum_{i \in S} q_i - \beta \sum_{i<j,\, i,j \in S} s_{ij}$ illustrates when retaining a second mode becomes preferable. Let the similarity between same-mode samples be ρ_{in} and cross-mode similarity be $\rho_{\text{out}} < \rho_{\text{in}}$. Replacing one member of an all-A set by a member of B changes the score by

$$F(S_{K-1,1}) - F(S_{K,0}) = -\alpha(q_A - q_B) + \beta(K-1)(\rho_{\text{in}} - \rho_{\text{out}}). \tag{16}$$

An all-A set is therefore suboptimal whenever a mode-B candidate is available and the pairwise gain exceeds the quality gap. This stylized model illustrates selection concentration under its score-ordering assumption and the population-dependent exchange margin introduced by pairwise interactions.

A.4 Fixed-cardinality solver

Let S be the active set. Constructive greedy begins at $S = \varnothing$ and adds

$$j^* = \arg\min_{j \notin S} \left(a_j + \sum_{i \in S} b_{\min(i,j),\max(i,j)} \right) \tag{17}$$

until $|S| = K$. One-swap refinement scans pairs (i, j) with $i \in S$ and $j \notin S$, applying the best energy-reducing pair found within a tolerance of 10^{-12}. The main dynamic selector stops after 100 scans or when a scan finds no reduction exceeding this tolerance. Candidate-index order resolves ties. Every refinement state and the returned selection satisfy $|S| = K$.

Implementation complexity. For M candidates with d-dimensional features and the self-reference choice $R = M$, forming the dense moment-kernel QUBO costs $O(M^2 d)$ time and $O(M^2 + Md)$ memory, including features. The released constructive solver recomputes the interaction sum over the selected columns at each addition: its selection loop costs $O(MK^2)$, with a further $O(M^2)$ cost to materialize the symmetric interaction matrix and evaluate the returned energy. In each one-swap scan, it enumerates each of the K possible removals and sums the remaining $K-1$ interactions for each unselected addition. For $1 \le K < M$, this costs $O(K^2(M-K+1))$ per scan. With $P \le 100$ scans, the combined objective construction and greedy-plus-swap computation therefore costs

$$O\big(M^2 d + MK^2 + PK^2(M-K+1)\big) \tag{18}$$

time and $O(M^2 + Md)$ memory. These bounds describe the implemented column-sum recomputation. At $K = M$, the entire population is retained and the refinement terminates without an exchange. Neural feature extraction and denoising are timed separately in the end-to-end profiles.

A.5 QUBO-to-Ising mapping

For hardware export, write $x_i = (s_i + 1)/2$, where $s_i \in \{-1, +1\}$. For

$$E(x) = \sum_i a_i x_i + \sum_{i<j} b_{ij} x_i x_j, \tag{19}$$

the nonconstant Ising Hamiltonian is

$$H(s) = \sum_i h_i s_i + \sum_{i<j} J_{ij} s_i s_j, \tag{20}$$

$$h_i = \frac{a_i}{2} + \frac{1}{4} \sum_{j \neq i} b_{\min(i,j),\max(i,j)}, \qquad J_{ij} = \frac{b_{ij}}{4}. \tag{21}$$

The exact-cardinality constraint becomes $\sum_i s_i = 2K - M$. Our physical-CIM package quantizes the complete signed Ising instance to int14 and adds one auxiliary spin for linear fields. The factor-of-two convention in the exported Kaiwu matrix is important. Introduce a gauge spin g and hardware spins $v = (s_1, \ldots, s_M, g)$. The logical spin is $\widetilde{s}_i = s_i g$, so $\widetilde{s}_i \widetilde{s}_j = s_i s_j$. Define a zero-diagonal symmetric matrix W by

$$W_{ij} = W_{ji} = J_{ij}/2, \qquad 1 \le i < j \le M, \tag{22}$$

$$W_{i,M+1} = W_{M+1,i} = h_i/2. \tag{23}$$

Because $v^\top W v$ counts every off-diagonal entry twice,

$$v^\top W v = \sum_{i<j} J_{ij} s_i s_j + \sum_i h_i s_i g = H(\widetilde{s}). \tag{24}$$

Consequently, a minimization backend receives W without a sign inversion, and a returned state is decoded as $x_i = (s_i g + 1)/2$. The global flip $v \mapsto -v$ leaves every logical bit and the energy unchanged.

For signed int14 export, let $\lambda = \max_{i<j} |W_{ij}|/8191$ and set $W_{ij}^{(14)} = \mathrm{round}(W_{ij}/\lambda)$. Up to a state-independent offset and coefficient-rounding residual,

$$E(x) = \lambda v^\top W^{(14)} v + C + \epsilon_{\mathrm{quant}}(x). \tag{25}$$

The integer Hamiltonian reported by the hardware is therefore $v^\top W^{(14)} v$ and lower is better. We audit both the direct integer energy and the decoded float64 QUBO energy because quantization can change the ordering of nearby states. A device minimizing $-v^\top J v$ instead receives $J = -W^{(14)}$; Appendix I audits the supplied physical returns under the original minimization Hamiltonian.

B Detailed Experimental Protocol

SiT time convention and stochastic branching. SiT uses the linear latent interpolant $z_t = t z_{\mathrm{clean}} + (1-t)\epsilon$, with data time $t = 0$ at Gaussian noise and $t = 1$ at the clean endpoint. The path-search noise level is $\sigma = 1 - t$. We first integrate $K = 50$ independent initial noises from $t = 0$ to $t = 0.89$, using 111 second-order Heun steps (222 model evaluations). Each retained parent at time t then produces B branches at the earlier time $u < t$ by

$$z_u = \frac{u}{t} z_t + \sqrt{(1-u)^2 - \left(\frac{u}{t}\right)^2 (1-t)^2}\, \epsilon', \qquad \epsilon' \sim \mathcal{N}(0, I), \tag{26}$$

using independent Gaussian noise for each branch. This is the Gaussian re-noising transition associated with the linear interpolant; the numerical implementation floors the expression under the square root at zero. The configured forward and backward increments in σ are $\Delta_f = 0.78$ and $\Delta_b = 0.81$. The backward endpoint is clipped at $\sigma = 0$, giving the four rounds in Table 5.

Table 5: Exact data-time schedule for the SiT dynamic search. Every backward leg uses five Heun steps (10 NFE per candidate). The first three rounds require one additional velocity evaluation per candidate for predicted-clean scoring.

Round	Parent t	Re-noised u	Backward endpoint	Extra scoring NFE
1	0.89	0.11	0.92	1
2	0.92	0.14	0.95	1
3	0.95	0.17	0.98	1
4	0.98	0.20	1.00	0

At the first three endpoints, a velocity prediction yields

$$\widehat{z}_{\mathrm{clean}}(z_t, t, \ell) = z_t + (1-t) v_\theta(z_t, t, \ell), \tag{27}$$

where ℓ is the requested class. Decoding this predicted-clean latent provides the images used for verifier scores and population features. The selected *intermediate states* z_t, rather than their predicted-clean counterparts, become the next round's parents. Round four already ends at $t = 1$: the selector scores and retains completed images, which are the final outputs. Thus the denoiser evaluation count per class and per retained output is, respectively,

$$\mathrm{NFE}_{\mathrm{class}} = 222K + KB(4 \cdot 10 + 3), \tag{28}$$

$$\mathrm{NFE}_{\mathrm{output}} = 222 + 43B. \tag{29}$$

This count includes the three predicted-clean evaluations and all branch legs. Feature extraction, decoding, and subset optimization contribute additional wall-clock cost, as measured in the full pipeline profiles.

Hyperparameter selection. We fixed $w_1 = w_2 = 1$ by construction in the principal SiT experiments. Before formal generation, a registered 16-class development screen compared constant $\alpha \in \{0.24, 0.40, 0.56\}$ and increasing/decreasing four-round schedules. The constant $\alpha = 0.56$ schedule was the only candidate that met the pre-specified requirements of at least 98.5% Top-K verifier-quality retention, at least 10% held-out moment reduction with a paired-bootstrap upper confidence bound below zero, and retention of at least 98% of the mean pairwise DINO distance relative to Top-K. Selection used these development criteria before formal ADM evaluation. The held-out moment reference comprises 64 separately generated Heun-250 images per development class at CFG 1.0; it is used for calibration, whereas online matching uses the current candidates. The same dynamic screen contained one DPP exponent, $\eta = 2$, so the dynamic tuning effort differed between methods. Subsequent matched-feature sweeps compare their quality–coverage operating points. A later dynamic calibration keeps this QUBO schedule fixed, screens DPP weights by measured ADM precision, and tests the selected weight on separate classes before extending a formal noise population. The complete sequence, including the earlier setting outside the matching margin, appears in Appendix C.14.

DiT single-pruning RBF control. We further test DiT-XL/2 (Peebles & Xie, 2023) on two independently seeded ImageNet-256 pools. DDIM (Song et al., 2021) uses 25 steps, classifier-free guidance 1.5, a 10-step prefix, $K = 50$, and $M = 200$, for 110 network-equivalent NFE per output. RBF-MMD-QUBO uses DINOv2 features and chooses the first α from a frozen grid that retains at least 99.5% of Top-K percentile quality. DPP uses Inception pre-logit features under the identical quality constraint.

Dynamic DPP quality weighting and greedy MAP inference. The dynamic DPP and QUBO share the raw target-class DINO logit q_i but transform it differently. QUBO uses the robust z-score in Appendix A in an additive quality term. DPP assigns ascending percentile rank $r_i \in \{0, \dots, M-1\}/(M-1)$, breaking tied logits by candidate index, and constructs

$$d_i = \exp\{\eta(r_i - \tfrac{1}{2})\}, \qquad L_{ij} = d_i \exp\left(-\frac{\|f_i - f_j\|_2^2}{2\sigma^2}\right) d_j. \tag{30}$$

The bandwidth σ is the square root of the median positive pairwise squared feature distance in the current candidate population, with fallback one when every distance vanishes. Features are DINO or Inception as specified for each comparison. We use $\bar{L} = L + 10^{-8} I$ and constructive greedy MAP inference: beginning at the empty set, repeatedly add the available candidate with largest $\log\det \bar{L}_{S\cup\{i\}} - \log\det \bar{L}_S$, using incremental Cholesky residuals floored at 10^{-8}, until $|S| = K$. Candidate-index order resolves ties. This is a greedy approximation to fixed-cardinality MAP. Thus the dynamic matched-feature comparison fixes the verifier and representation, while quality transformations and objective forms differ; the later calibration also matches observed precision.

The frozen-pool quality-constrained DPP experiments use a separate rule: an unweighted RBF kernel, jitter 10^{-6}, and log-determinant-improving one-out/one-in refinement from Top-K under the stated quality floor, with a 100-scan cap. This hard constraint differs from dynamic exponential quality weighting.

Additional search and selection context. Existing approaches change exploration through local/global search (Zhang et al., 2026) and evolutionary mutations (He et al., 2025), or improve intermediate values through terminal rewards, lookahead and calibration (Jain et al., 2025; Oshima et al., 2025). Population control includes potential-based particle steering (Singhal et al., 2025), partitioned DPP beams in discrete text diffusion (Lys et al., 2026), progressive pruning of initial seeds (Guimarães & Perona, 2026), and variance-aware resampling (Shekhar et al., 2026). Particle Guidance and MMD Guidance alter diffusion dynamics through joint potentials and reference-matching gradients (Corso et al., 2024; Mahdizadeh Sani et al., 2026). Related work formulates continuous distributional control through a mean-field framework (Howard & Nüsken, 2026).

Scalable Group Inference (Parmar et al., 2026) provides a closely related discrete formulation: it combines unary quality and pairwise diversity in a fixed-cardinality quadratic integer program, solved with Gurobi branch-and-cut. Its default scores use CLIP quality and DINO cosine distance. Intermediate clean predictions guide progressive pruning of an initial pool of M trajectories until K remain. Thus, quadratic subset selection and intermediate pruning are shared with our approach.

Table 6: Identification logic of the principal controls. The comparisons separate the effects of population selection, feature representation, pairwise structure, and solver depth.

Comparison	Held fixed	Question isolated
Top-K vs. population QUBO	candidate generator, NFE, verifier, K	Does set-aware selection mitigate collapse induced by independent ranking?
DINO-DPP vs. Moment-QUBO	candidate budget and output size	Does QUBO beat a strong set-aware baseline when representation and objective are both allowed to differ?
Inception-DPP vs. Moment-QUBO	Inception representation, DINO verifier, K, NFE	What operating point is induced by determinant diversity versus moment matching?
Full vs. linear vs. shuffled QUBO	frozen 200K pool, linear scores, solver	Do pairwise terms help, and does the identity of each coupling matter?
Constructive vs. one-swap/FSA	identical QUBO instance	How much of the result is objective design versus stronger optimization?
DINO-KID and CLIP-CMMD audits	completed generated images	Does the measured advantage survive a change of evaluation representation?

Our multi-round setting repeatedly expands K retained parents into KB new branches and selects K parents again, so retention changes the candidates explored in subsequent rounds. Moment-QUBO also assigns an explicit statistical target to each decision: the current candidates supply the reference moments against which the selected subset is matched. This lets us study the distributional distortion of pruning within a repeated branching process.

Noise-space optimization studies the starting state of a diffusion trajectory (Qi et al., 2024). MMR and k-medoids provide classical diversity-aware subset rules (Carbonell & Goldstein, 1998; Kaufman & Rousseeuw, 1990). The comparisons below focus on the effects of retaining a population, matching its feature representation, and optimizing its explicit objective.

Table 7: SiT dynamic path-search protocol.

Component	Setting
Backbone	SiT-XL/2, ImageNet-256
Sampler	deterministic second-order Heun
Classifier-free guidance	1.0
Initial population	$K = 50$ paths per class
Search rounds	$T = 4$
Branches per path	$B \in \{1, 2, 3, 4\}$
Candidates per round	$M = KB \in \{50, 100, 150, 200\}$
Path segment	10 NFE
Initial integration	222 NFE
Total per output	$222 + 43B$ NFE
Quality verifier	target-class DINOv2 ViT-L/14 logit
QUBO feature	Inception-v3 pool3, 2048 dimensions
QUBO objective	self-reference moment MMD + quality
QUBO weights	$\alpha = 0.56$, $w_1 = w_2 = 1$
Solver	greedy-K + best one-swap descent
DPP baseline	DINO RBF, quality exponent $\eta = 2$
Outputs	50/class, 1,000 classes

The initial-noise seed is a deterministic function of class and within-class index. Branch seeds use a disjoint namespace and the index $r \cdot 10000 + p \cdot 4 + b$, where r, p, b are zero-based round, parent-slot and branch indices. Thus branch-noise indices are nested across budgets for the same parent slot. Once selected parents differ, later candidate sets can diverge across budgets. Every budget executes all four rounds; only branches per parent change. At $B = 1$, each round has exactly 50 candidates and all selectors must retain the complete set.

For this RBF control, α is chosen per class from the frozen grid

$$\{0.10, 0.12, 0.15, 0.18, 0.20, 0.25, 0.30, 0.40, 0.60, 0.80, 1.00\}$$

Table 8: DiT single-pruning RBF-control protocol.

Component	Setting
Backbone	DiT-XL/2, ImageNet-256
Sampler	deterministic DDIM, 25 steps
Classifier-free guidance	1.5
Prefix	10 steps
Population/candidates	$K = 50$, $M = 200$
Total per output	110 network-equivalent NFE
Quality	target-class DINOv2 linear-head logit
QUBO feature/kernel	DINOv2 class token / median RBF
QUBO reference	all 200 same-class candidates
Quality rule	at least 0.995 of Top-K percentile quality
DPP feature	Inception-v3 pre-logit
Replicate seeds	independent namespaces 480812 and 580812
Outputs	50/class, 1,000 classes, per replicate

as the first setting that reaches the frozen quality threshold. Both DPP and RBF-MMD-QUBO use the same candidate pool and quality threshold. A cache is allowed only for candidate indices already generated under the same deterministic suffix; an audit confirmed pixel-exact regeneration for identical indices.

C Complete Formal Results

C.1 Pure sampling and the no-pruning path configuration

The original formal evaluation also includes pure Heun-250 with the same SiT-XL/2 checkpoint, guidance 1.0, and 50 outputs per class. This baseline samples without path branching or intermediate selection. Its higher recall and lower precision locate it at another quality–coverage operating point. The $B = 1$ configuration in the search curve instead executes all four path rounds at 265 NFE, retaining all 50 candidates each time.

Table 9: Pure-sampler reference from the original official ImageNet-256 50K evaluation. This configuration differs from the 265-NFE $B = 1$ path curve.

Method	NFE	FID	sFID	IS	Prec.	Rec.
Pure Heun-250	250	9.35168	6.67874	43.48605	0.66504	0.67170

C.2 Retention diagnostics in the principal $K = 50$ setting

We reanalyze saved candidate features, quality scores, selection indices, and parent identities from the two principal SiT noise populations. The analysis covers 16,000 class/method/budget archives and 64,000 round records, with 1,000 classes per available configuration. Source feature hashes and exact cardinality are checked before analysis. Root entropy is normalized by $\log 50$; parent fraction counts the distinct immediate parents retained, divided by 50. Pairwise DINO distance uses the unnormalized candidate features stored during generation.

Figure 6 follows the actual populations at $B = 4$. Table 10 gives final-round Moment-QUBO-minus-Top-K differences and 2,000-replicate paired class-bootstrap intervals. These comparisons use each policy's actual candidate population; later pools differ because earlier selected parents differ. The increased lineage and feature diversity therefore describe complete policy trajectories.

An additional conditional diagnostic applies Top-K to the very same candidate pool and ancestry produced by each actual policy, without rolling the counterfactual selection forward. On Moment-QUBO's final $B = 4$ pools, its selected-to-candidate moment discrepancy is lower than this own-pool Top-K by 0.05908/0.05832 (original/fresh); paired intervals are $[-0.06194, -0.05612]$ and $[-0.06141, -0.05510]$. This isolates an immediate pruning effect conditional on the existing candidate pool. The moment kernel, reference center, and scales are fixed within each such comparison.

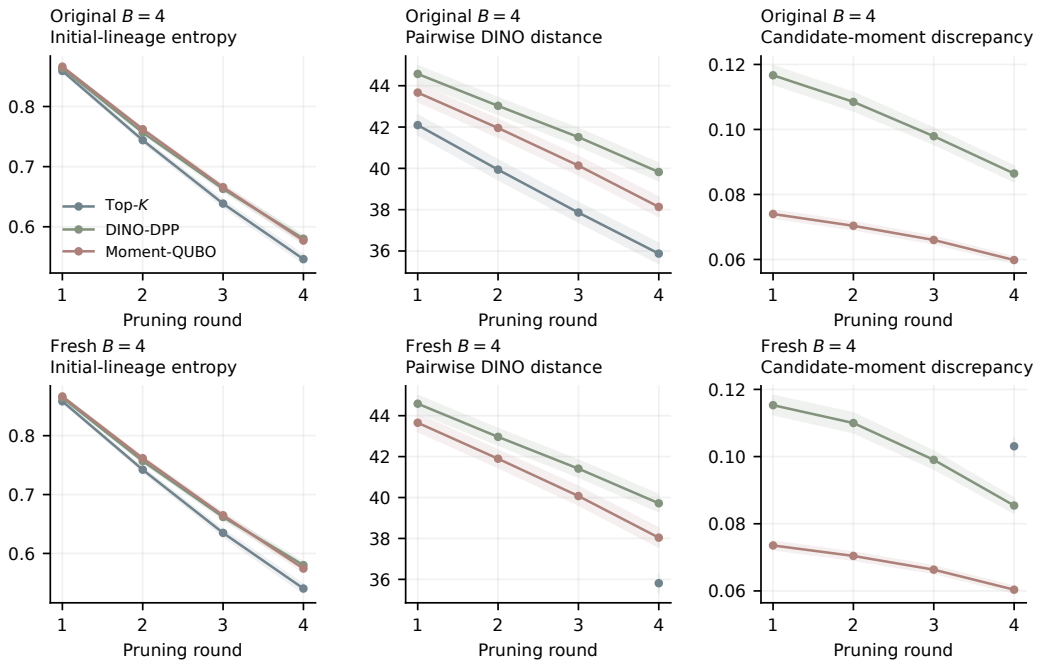

Figure 6: Principal $K = 50$, $B = 4$ retention diagnostics, original and fresh noise populations. Bands show pointwise class-bootstrap 95% intervals. All methods lose initial-lineage diversity over rounds; population-aware selectors retain more. Moment discrepancy is computed relative to each policy's own candidate pool with that pool's normalization. Original Top-K has no roundwise Inception cache; fresh Top-K has feature caches only for the final round. Missing points are omitted.

Table 10: Final-round Moment-QUBO minus Top-K on each full class set. Each interval pairs the same 1,000 classes within one noise replication.

Noise	B	Entropy delta	95% CI	DINO distance delta	95% CI
Original	2	$+0.0108$	$[+0.0082, +0.0132]$	$+1.0404$	$[+0.9340, +1.1499]$
Original	3	$+0.0219$	$[+0.0183, +0.0257]$	$+1.8576$	$[+1.7244, +1.9958]$
Original	4	$+0.0309$	$[+0.0260, +0.0356]$	$+2.2600$	$[+2.1017, +2.4179]$
Fresh	2	$+0.0085$	$[+0.0061, +0.0110]$	$+1.1823$	$[+1.0702, +1.2987]$
Fresh	3	$+0.0255$	$[+0.0218, +0.0292]$	$+1.8656$	$[+1.7200, +2.0128]$
Fresh	4	$+0.0339$	$[+0.0289, +0.0390]$	$+2.2263$	$[+2.0834, +2.3676]$

C.3 Main-objective intervention at the final pruning round

To distinguish final image selection from earlier changes to the retained parents, we compare three complete policies on the fresh $B = 4$ noise population. T denotes stable pointwise Top-K; Q denotes the main self-reference moment objective with $\alpha = 0.56$, $w_1 = w_2 = 1$, and constructive greedy followed by one-swap refinement. TTTT and QQQQ are the existing four-round policies. TTTQ reuses TTTT's first three rounds and its complete final candidate pool, changing only the final selection rule.

The 200 candidates per class are already decoded at the final zero-noise endpoint, so the intervention adds zero denoising NFE. All 1,000 classes pass stable Top-K replay, and replaying Q on the archived QQQQ final pools exactly reproduces all recorded QQQQ selections. Each selected image is verified against its candidate hash before assembling the class-balanced 50K evaluation archive. The intervention changes 6,308 of Top-K's 50,000 selected images. Its mean per-class DINO-quality retention is 0.995002, an outcome of the weighted objective rather than an imposed quality floor.

Table 11: Retention-policy intervention with the main $\alpha = 0.56$ objective. All rows use the original CPU ADM evaluation environment on 50K images. The three policies share initial states and branch-noise namespaces; TTTT and TTTQ additionally share every candidate pool.

Policy	FID	sFID	IS	Prec.	Rec.
TTTT	9.52512	9.17054	67.32794	0.88308	0.32850
TTTQ	7.76303	9.01483	65.62344	0.86106	0.37320
QQQQ	6.45538	8.52897	64.26546	0.84292	0.41190

This post-hoc intervention uses an already evaluated independent noise pool. The policy contrast measures the effect of applying joint retention earlier in the search; the intervening candidate populations then evolve differently. FID differences are reported between complete policies. A separate complete GPU evaluation reproduces the two existing policies' FID within 4.7×10^{-7}, while their precision/recall values differ by up to 0.00064 from CPU values, exceeding the predeclared 0.0002 P/R reproduction tolerance. Table 11 retains the common CPU backend.

C.4 Fixed-pool final-pruning controls under a quality floor

On a separate noise population (initial/branch namespaces 48/49), a post-hoc class-balanced 50K control study holds the first three Top-K rounds and all 200 final candidates per class fixed, then changes only the last selection. Here Moment-QUBO minimizes self-reference moment discrepancy subject to retaining at least 0.995 of Top-K mean DINO quality. This hard quality-floor objective differs from the main dynamic $\alpha = 0.56$ weighted objective. All selectors return exactly 50 images per class; the unchanged generation protocol costs 394 logical denoising NFE per output.

Table 12: Shared-final-pool controls: official ImageNet-256 50K metrics. The quality-floor selectors satisfy a common DINO-score floor; their actual quality and ADM precision can differ. Parent-stratified Top-K is shown as a quality-unmatched ancestry control.

Selector	FID	sFID	IS	Prec.	Rec.
Top-K	9.67073	8.96080	67.57033	0.88600	0.32380
Moment-QUBO	8.35647	8.87164	66.40908	0.87106	0.35750
Moment-greedy	8.74195	8.89359	66.89499	0.87700	0.35150
Linear-only	9.77898	8.97679	67.63957	0.88948	0.32950
Shuffled quadratic	9.61089	8.96713	67.62715	0.88558	0.32890
Quality-floor parent coverage	9.17503	8.80190	67.32259	0.88146	0.34830
Quality-floor root coverage	9.26233	8.79495	67.33250	0.88344	0.34750
Parent-stratified Top-K (unmatched)	7.07032	8.46354	64.90072	0.86012	0.41010
Inception-DPP	8.18416	8.81875	67.02272	0.86056	0.37420

Within this protocol, refinement improves FID by 0.38548 relative to same-objective constructive greedy. Linear-only and identity-shuffled quadratic controls have 1.42252 and 1.25442 higher FID. These are different operating conditions from both the main weighted dynamic objective and the global real-reference structure test in Table 36. Inception-DPP attains lower FID and higher recall than Moment-QUBO, with lower precision. The unconstrained parent-stratified rule also has lower FID, while retaining approximately 0.9615 of Top-K mean DINO quality in the cached diagnostic.

Ancestry and feature coverage separate even among the quality-constrained controls. On the frozen 800-class cached validation split, parent/root coverage heuristics have root entropies 0.5784/0.5867, compared with 0.5523 for Moment-QUBO. Their formal 50K FID is higher and recall lower. Lineage retention is thus a useful process diagnostic alongside the statistics of the retained feature population.

Table 13: SiT-XL/2 full four-round Search-over-Paths results. Each row is an official class-balanced ImageNet-256 50K evaluation. Bold marks the best value within each nontrivial budget; metrics at $B = 1$ are identical.

B	NFE	Method	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
1	265	All selectors	16.590	12.804	39.021	0.6182	0.6543
2	308	Top-K	5.456	**8.387**	**61.247**	**0.8352**	0.4574
2	308	DINO-DPP	4.977	8.426	59.984	0.8107	**0.4972**
2	308	Moment-QUBO	**4.851**	8.484	59.752	0.8144	0.4914
3	351	Top-K	7.999	8.841	**65.782**	**0.8726**	0.3713
3	351	DINO-DPP	6.318	8.611	64.071	0.8432	**0.4349**
3	351	Moment-QUBO	**5.815**	**8.522**	63.085	0.8358	0.4320
4	394	Top-K	9.556	9.060	**67.562**	**0.8829**	0.3302
4	394	DINO-DPP	7.114	8.634	65.465	0.8465	**0.4061**
4	394	Moment-QUBO	**6.423**	**8.545**	64.494	0.8474	0.4026

Table 14: Moment-QUBO differences from the two SiT controls. Negative FID and positive recall changes are favorable.

	QUBO − Top-K		QUBO − DINO-DPP	
B	FID	Recall	FID	Recall
2	-0.6046	+0.0340	-0.1264	-0.0058
3	-2.1835	+0.0607	-0.5021	-0.0029
4	-3.1334	+0.0724	-0.6918	-0.0035

C.5 Complete SiT metrics

C.6 SiT deltas by budget

The $B = 4$ result was known before the complete curve was registered. The $B = 1, 2, 3$ image sets and ADM metrics were generated afterward. The four budgets form one paired compute curve with shared initial states and nested branch-noise indices; the following study supplies an independent noise replication.

C.7 Independent fresh-noise scaling replication

The prospectively registered independent run uses initial- and branch-noise namespaces 78 and 79, disjoint from the original curve, while freezing the SiT checkpoint, four-round schedule, $K = 50$, $B = 4$, objective weights, selectors, and 394 NFE/output. Both endpoint gates—lower QUBO FID and higher QUBO recall than Top-K—passed. Once that result was known, a separate strengthening protocol froze $B = 1, 2, 3$ before any corresponding image or ADM metric existed. It reused only the independent initial states and nested branch-noise construction. Every method has 50 images per ImageNet class.

The registered conditions all pass: QUBO improves FID and recall at each nontrivial budget, and its absolute FID advantage is non-decreasing from $B = 2$ through $B = 4$. The full metric table, including the $B = 4$ DINO-DPP endpoint, is Table 15.

C.8 DiT nested-prefix static noise-search curve

We also evaluate complete static candidate-search curves on DiT-XL/2. For each class, budget B exposes the first $50B$ independently seeded, fully denoised candidates from each frozen 200-candidate pool and retains exactly 50. The candidate prefixes are nested, so the network-equivalent cost is $50B$ NFE/output under 25-step classifier-free-guided DDIM. The $B = 1$ and $B = 4$ endpoints in both pools were known before their respective curve-completion studies were registered; the $B = 2, 3$ selections and ADM metrics were outcome-blind. The two pools use disjoint candidate-

Table 15: Independent fresh-noise SiT scaling replication. Every row is a class-balanced ImageNet-256 50K evaluation. $B = 4$ was completed first; $B = 1, 2, 3$ were then prospectively registered before generation.

B	NFE	Method	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
1	265	All selectors	16.571	12.750	38.972	0.6170	0.6609
2	308	Top-K	5.545	**8.505**	**61.253**	**0.8344**	0.4501
2	308	Moment-QUBO	**4.892**	8.576	59.268	0.8120	**0.4898**
3	351	Top-K	7.935	8.725	**65.785**	**0.8731**	0.3745
3	351	Moment-QUBO	**5.880**	**8.502**	62.959	0.8365	**0.4351**
4	394	Top-K	9.525	9.171	**67.328**	**0.8831**	0.3285
4	394	DINO-DPP	7.172	8.579	65.648	0.8484	0.4089
4	394	Moment-QUBO	**6.455**	**8.529**	64.265	0.8429	**0.4119**

Table 16: Registered QUBO differences in the independent curve. Negative FID and positive recall changes are favorable.

B	FID: QUBO − Top-K	Recall: QUBO − Top-K
2	-0.653287	+0.039700
3	-2.055813	+0.060600
4	-3.069747	+0.083400

seed namespaces. This post-hoc use of frozen pools tests static noise selection under increasing candidate compute; the SiT curves evaluate multi-round path search.

The DiT curves use a global multi-feature QUBO variant, distinct from the class-local self-reference objective in the SiT curves. Let S contain $N = 50{,}000$ selected images, S_c its $K = 50$ members in class c, and $\mathcal{H}_b(S)$ the normalized first/second-moment discrepancy in Equation 45 for feature block b, using that block's real-reference mean, covariance, and normalization. With DINO features g_i, the minimized objective is

$$E_{\mathrm{DiT}}(S) = \mathcal{H}_{\mathrm{Inc}}(S) + 0.5\mathcal{H}_{\mathrm{DINO}}(S) - 0.05\frac{V(S)}{V(S_{\mathrm{DPP}})}, \tag{31}$$

$$V(S) = \frac{1}{1000}\sum_{c=1}^{1000}\frac{2}{K(K-1)}\sum_{i<j,\, i,j\in S_c}\|g_i - g_j\|_2^2. \tag{32}$$

Both moment weights within each block are one. Inception has 2048 dimensions and DINO has 1024. The real-reference statistics are fixed across budgets; the dispersion normalization is fixed at the DPP initialization for each pool–budget problem. The global objective remains quadratic at fixed class cardinalities. It is minimized subject to 50 images per class and a per-class mean raw DINO logit at least 0.995 of the corresponding Top-K mean. Starting from quality-constrained Inception-DPP, an implicit global solver scans all feasible same-class one-out/one-in moves and applies one best decreasing move per scan, capped at 1000 scans with tolerance 10^{-14} and FP32 arithmetic. The DPP initializer uses the hard-floor rule in Appendix B. These curves therefore test a real-reference, multi-feature static selection variant; its small increments over DPP include refinement of that shared starting selection.

Across the two pools, the mean Top-K–QUBO FID gaps are 1.0578, 1.5124, and 1.7810 at $B = 2, 3, 4$, with respective between-pool ranges $[1.0416, 1.0740]$, $[1.5103, 1.5145]$, and $[1.7467, 1.8154]$. The corresponding mean QUBO recall gains are 0.0425, 0.0579, and 0.0625. The FID gap is non-decreasing over all four nested budgets within each pool. Inception-DPP is also population-aware and closely follows QUBO: QUBO lowers FID by 0.0281–0.1006 across individual nontrivial pool–budget blocks, with mixed and small precision/recall differences. Both population selectors preserve the widening FID advantage over Top-K as the candidate budget grows.

Table 17: Two independent DiT-XL/2 nested-prefix static noise-search pools. Here Moment-QUBO denotes the global multi-feature variant defined above. Each reported row is a class-balanced ImageNet-256 50K ADM evaluation. At $B = 1$, all selectors are identical. Bold marks the best metric within each nontrivial pool–budget block.

Pool	B	NFE	Selector	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
A	1	50	Top-K	3.015	4.769	57.382	0.7892	0.5755
	2	100	Top-K	5.047	5.149	**66.109**	**0.8767**	0.4603
	2	100	Inception-DPP	4.001	**5.058**	65.558	0.8542	**0.5030**
	2	100	Moment-QUBO	**3.973**	5.073	65.503	0.8550	0.4996
	3	150	Top-K	6.385	5.421	**67.699**	**0.8891**	0.4295
	3	150	Inception-DPP	4.918	**5.282**	67.154	0.8635	0.4874
	3	150	Moment-QUBO	**4.870**	5.289	67.090	0.8641	**0.4883**
	4	200	Top-K	7.257	5.638	**68.536**	**0.8953**	0.4066
	4	200	Inception-DPP	5.542	5.491	68.136	0.8687	0.4653
	4	200	Moment-QUBO	**5.441**	**5.486**	68.088	0.8688	**0.4654**
B	1	50	Top-K	3.050	4.739	57.765	0.7905	0.5801
	2	100	Top-K	4.987	5.159	**65.976**	**0.8784**	0.4537
	2	100	Inception-DPP	3.988	**5.056**	65.408	0.8557	**0.4993**
	2	100	Moment-QUBO	**3.945**	5.061	65.337	0.8564	**0.4993**
	3	150	Top-K	6.315	5.430	**67.776**	**0.8915**	0.4271
	3	150	Inception-DPP	4.876	5.285	67.270	0.8665	0.4819
	3	150	Moment-QUBO	**4.805**	**5.275**	67.299	0.8668	**0.4840**
	4	200	Top-K	7.138	5.651	**68.654**	**0.8972**	0.4114
	4	200	Inception-DPP	5.452	**5.469**	68.157	0.8713	0.4754
	4	200	Moment-QUBO	**5.391**	5.476	68.103	0.8720	**0.4775**

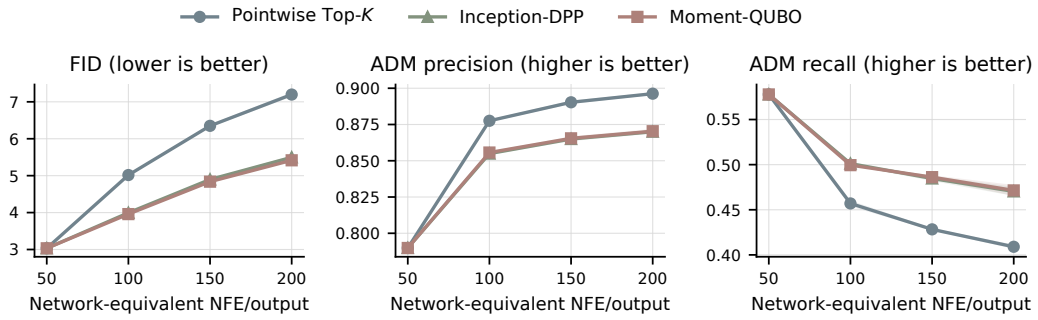

Figure 7: Formal DiT static noise-search compute curves. Lines are the two-pool means and bands span the two independent pools. Within each pool, the $B = 1$ point is shared because candidate and retained populations both contain 50 samples per class.

C.9 Paired uncertainty for official ADM FID

For completed 50K archives, we extract the exact official ADM pool3 and spatial activations and compute a paired class-cluster influence-function bootstrap with 20,000 resamples. Each resample applies the same ImageNet-class weights to both methods, and the interval is centered on the exact official metric difference. Table 18 distinguishes large population effects from smaller structure effects.

These intervals are conditional on each frozen archive and the fixed ADM reference and quantify sensitivity across ImageNet classes. Independently regenerated noise populations are evaluated in

Table 18: Paired ADM uncertainty. Deltas are first method minus second method; negative values favor the first method.

Comparison	FID Δ	95% CI	sFID Δ	95% CI
SiT dynamic QUBO vs. Top-K	-3.1334	$[-3.6743, -2.6202]$	-0.5149	$[-0.7214, -0.3114]$
SiT dynamic QUBO vs. DINO-DPP	-0.6918	$[-1.1432, -0.3110]$	-0.0887	$[-0.2549, +0.0811]$
Fresh full QUBO vs. linear-only	-0.0638	$[-0.1215, -0.0197]$	-0.0090	$[-0.0189, +0.0008]$
Fresh full QUBO vs. shuffled	-0.0581	$[-0.0845, -0.0323]$	-0.0165	$[-0.0283, -0.0048]$
DiT QUBO vs. Top-K	-1.8154	$[-1.9637, -1.6639]$	-0.1518	$[-0.1944, -0.1090]$
DiT QUBO vs. DPP	-0.1006	$[-0.1386, -0.0651]$	-0.0049	$[-0.0198, +0.0098]$

Table 19: Conditional paired query-resampling intervals for the six listed comparisons; differences are first method minus second. The SiT dynamic DPP comparator is DINO-DPP. DiT uses static pool A at $B = 4$ in Table 17, with Moment-QUBO and Inception-DPP. The calibrated RBF-MMD results in Table 37 use separate candidate pools.

Comparison	Δ precision	Conditional 95% interval	Δ recall	Conditional 95% interval
SiT dynamic QUBO − Top-K	-0.03554	$[-0.04000, -0.03106]$	$+0.07240$	$[+0.06350, +0.08160]$
SiT dynamic QUBO − DINO-DPP	$+0.00088$	$[-0.00392, +0.00578]$	-0.00350	$[-0.01260, +0.00570]$
Fresh full QUBO − linear-only	-0.00082	$[-0.00124, -0.00042]$	$+0.00020$	$[-0.00120, +0.00160]$
Fresh full QUBO − shuffled	-0.00068	$[-0.00112, -0.00026]$	$+0.00020$	$[-0.00150, +0.00190]$
DiT static Moment-QUBO − Top-K	-0.02658	$[-0.02798, -0.02518]$	$+0.05880$	$[+0.05350, +0.06400]$
DiT static Moment-QUBO − DPP	$+0.00006$	$[-0.00058, +0.00072]$	$+0.00010$	$[-0.00280, +0.00310]$

the replication studies. The DiT rows correspond to the historical static moment-selector endpoint; Table 37 reports the two calibrated RBF-MMD endpoints. IS remains a point estimate; the following analysis adds conditional P/R intervals for these same six registered archive comparisons.

C.10 Conditional paired precision and recall intervals

We supplement the six archive comparisons in Table 18 with paired uncertainty for ADM precision and recall (Kynkäänniemi et al., 2019; Dhariwal & Nichol, 2021). This analysis reuses 9 frozen 50,000-image generated archives and the same 10,000 real reference images. Let $\mathcal{M}_R$ denote the fixed real manifold and $\mathcal{M}_a$ the fixed generated manifold of method a. We retain the official query membership indicators and calculate

$$P_a = \frac{1}{50000} \sum_i \mathbf{1}[f_i^a \in \mathcal{M}_R], \qquad R_a = \frac{1}{10000} \sum_j \mathbf{1}[y_j \in \mathcal{M}_a]. \tag{33}$$

Precision resamples the 1,000 verified ImageNet classes, retaining all 50 generated queries within each sampled class. Recall resamples the same real-query identities across the two methods. The reference archive provides unlabeled images, so recall uses image query units and assumes their exchangeability. For each comparison we use 20,000 paired percentile bootstrap draws and report 95% intervals for the mean membership difference. The two methods always use the same resampled class/query indices.

Conditional scope and replay. The intervals measure query-composition sensitivity conditional on the frozen generated archives, hyperparameters, manifold centers, and nearest-neighbor radii. We report unadjusted per-comparison percentile intervals using the unchanged official evaluator's $k = 3$ neighborhood, 10,000-by-10,000 distance blocks, and float16 distance arithmetic with its original float32 fallback. Historical CPU execution exactly reproduces every archived precision and recall point estimate before resampling. These verified CPU membership indicators define the intervals; GPU replay differences are archived as numerical diagnostics.

C.11 Verifier-quality floor and ADM precision

We additionally complete official ADM evaluation of the independently frozen $B = 4$ final candidate pool. The global Moment-QUBO and Inception-DPP both keep exactly 50 candidates per class

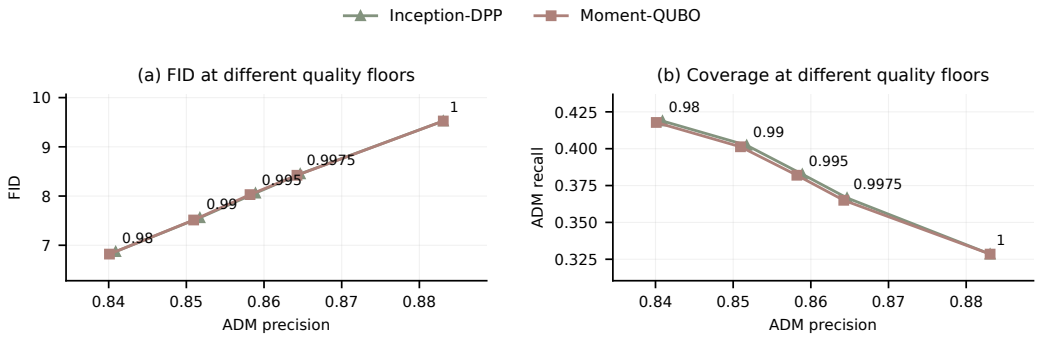

Figure 8: Global real-reference Moment-QUBO versus DPP on a frozen final pool. Each marker selects and evaluates 50K images at one DINO-quality floor; lines interpolate between measured configurations. This post-hoc selection sweep adds zero diffusion NFE. Dynamic operating points are reported separately in Tables 22 and 1.

subject to at least 0.995 of Top-K mean DINO quality. Their selected indices were frozen before this evaluation, and all images are reused from the 200K pool, giving zero additional diffusion NFE.

Table 20: Independent 50K final-pool selection under the 0.995 DINO-quality floor. Bold marks the best value in each column.

Selector	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
Pointwise Top-K	9.5251	9.1705	**67.3279**	**0.8831**	0.3285
Inception-DPP	8.0593	9.0029	66.7215	0.8589	**0.3827**
Global Moment-QUBO	**8.0293**	**9.0018**	66.6702	0.8582	0.3820

Global Moment-QUBO improves FID by 1.4959 and recall by 0.0535 over Top-K, but loses 0.0249 precision and therefore fails the registered 0.02 precision non-inferiority gate. Against matched Inception-DPP, its FID is lower by 0.0300, while precision and recall are each lower by 0.0007. The measured operating points retain a quality–coverage trade-off under the common verifier-quality floor.

C.12 Matched-feature final-pool Pareto sweep

Table 21 reports every point from the post-hoc quality-floor sensitivity study on the same frozen 200K candidate pool. The $q = 0.995$ indices exactly reproduce the independently frozen formal selection; at $q = 1$, both methods exactly equal Top-K. The other six archives are new class-balanced 50K evaluations. The frozen OpenAI evaluator's legacy `np.bool` alias is restored through the registered compatibility shim without modifying the evaluator; the first two attempts that reached FID/sFID/IS match the compatibility rerun exactly.

Figure 8 plots the measured configurations and the interpolated curves. The detailed matched-precision calculation below uses the same frozen-pool sweep.

At the same $q < 1$, QUBO improves FID by 0.028–0.050 but gives up at most 0.0008 precision and 0.0016 recall. Because even those small precision shifts can explain the FID change, we piecewise-linearly interpolate the DPP curve at each QUBO precision. For $q = 0.99, 0.995, 0.9975$, QUBO minus interpolated DPP is $+0.0049, +0.0189, +0.0002$ FID and $-0.0022, -0.0026, -0.0028$ recall, respectively. This descriptive matched-precision analysis indicates nearly overlapping set-aware frontiers on this pool.

Table 21: Global real-reference Inception-feature DPP/Moment-QUBO sweep. Both methods obey the same per-class DINO-quality floor and select exactly 50 of 200 candidates per class.

q	Selector	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
0.9800	Inception-DPP	6.8709	8.8726	66.3743	0.8409	0.4188
	Moment-QUBO	6.8205	8.8612	66.2975	0.8401	0.4178
0.9900	Inception-DPP	7.5592	8.9534	66.4778	0.8517	0.4023
	Moment-QUBO	7.5133	8.9427	66.4352	0.8509	0.4013
0.9950	Inception-DPP	8.0593	9.0029	66.7215	0.8589	0.3827
	Moment-QUBO	8.0293	9.0018	66.6702	0.8582	0.3820
0.9975	Inception-DPP	8.4506	9.0364	66.9558	0.8647	0.3666
	Moment-QUBO	8.4222	9.0405	66.9148	0.8642	0.3650
1.0000	Inception-DPP	9.5251	9.1705	67.3279	0.8831	0.3285
	Moment-QUBO	9.5251	9.1705	67.3279	0.8831	0.3285

Table 22: Dynamic matched-feature controls on SiT-original; Inception-DPP uses $\eta = 2$. All entries are class-balanced ImageNet-256 50K evaluations.

B	Selector	FID↓	Prec.↑	Rec.↑
2	Inception-DPP	**4.5514**	0.7648	**0.5480**
	Moment-greedy	4.8553	0.8141	0.4869
	Moment-one-swap	4.8510	**0.8144**	0.4914
4	Inception-DPP	**5.4513**	0.7770	**0.4916**
	Moment-greedy	6.4643	0.8461	0.4002
	Moment-one-swap	6.4227	**0.8474**	0.4026

C.13 Matched-feature operating points

Table 22 holds the population representation fixed. The lower-precision Inception-DPP points have lower FID and higher recall; the following calibration therefore compares objectives near the same precision rather than treating these operating points as interchangeable.

C.14 Dynamic calibration and a same-noise 50K extension

The frozen-final-pool sweep changes only which decoded images are retained. Here each tested selector instead makes all four dynamic pruning decisions, so its retained parents determine its later candidate population. All runs use $K = 50$, $B = 4$, the DINO class verifier, and Inception population features. QUBO keeps $\alpha = 0.56$, $w_1 = w_2 = 1$, and the original greedy-plus-swap solver. DPP uses the same frozen greedy MAP implementation while varying its quality exponent η.

Initial development and the first formal comparison. A 64-class screen first measured three QUBO weights and three DPP exponents on shared initial states, with 50 outputs per class (Table 23). A subsequent $\eta = 6$ measurement gave precision 0.8634375. Among the measured $\eta = 4, 6, 8$ candidates, $\eta = 8$ was nearest to QUBO precision 0.850000, with a gap of 0.007500. The then-registered rule promoted the nearest point even outside the 0.005 matching margin. The resulting formal experiment used fresh initial/branch namespaces 9140/9141, froze the setting before generation, and generated 50K images per method. Its observed precision gap was 0.01958, outside the matching margin. Table 25 retains this endpoint alongside the subsequent extension.

Two separate class subsets before another formal endpoint. After observing the first formal comparison, we registered a stricter calibration. A deterministic split (seed 2026091517) selects 128

Table 23: Initial dynamic development measurements: 64 classes, 3,200 images per setting. FID and recall use the complete real reference against a class-restricted generated population; their absolute values have a different sample-count and class-coverage basis from the formal 50K results. The follow-up setting was measured after the six-point screen.

Setting	Development FID↓	Precision↑	Recall↑
Moment-QUBO ($\alpha = 0.28$)	59.28518	0.809375	0.0870
Moment-QUBO ($\alpha = 0.56$)	66.67749	0.850000	0.0608
Moment-QUBO ($\alpha = 1.12$)	71.71852	0.862187	0.0466
DPP ($\eta = 2$)	62.67928	0.790312	0.0955
DPP ($\eta = 4$)	70.52817	0.838437	0.0683
DPP ($\eta = 8$)	74.63293	0.857500	0.0479
DPP ($\eta = 6$, follow-up)	73.45978	0.863437	0.0624

screening and 128 held-out classes, mutually disjoint and excluding the previous 64 development classes. Both subsets use the existing 78/79 noise namespaces, with fixed QUBO outputs reused. The screen measures DPP $\eta = 4, 5, 6$ and selects the smallest absolute precision gap, breaking ties by smaller η. It must satisfy $|P_Q - P_D| \leq 0.005$ before the selected setting alone is tested on the held-out subset. Failure on either subset stops formal extension; FID is absent from the selection criterion.

Table 24: All settings from the stricter calibration. Each method/setting uses 6,400 images per subset. Differences are QUBO minus DPP. Both gates use the observed point difference; conditional intervals are reported separately in the text.

Subset	η	QUBO precision	DPP precision	Difference	Gate
Screen (128 classes)	4	0.836562	0.831562	+0.005000	Pass
Screen (128 classes)	5	0.836562	0.847812	-0.011250	Outside margin
Screen (128 classes)	6	0.836562	0.851094	-0.014531	Outside margin
Held-out (128 classes)	4	0.847969	0.846875	+0.001094	Pass

DPP $\eta = 4$ passes both point gates: the screen gap is exactly 0.005000, and the held-out gap is 0.001094. Their fixed-reference paired class-query 95% intervals are $[-0.00969, 0.01953]$ and $[-0.01219, 0.01453]$. The gate therefore selects similar measured precision; the wider intervals quantify the uncertainty associated with these 128-class subsets.

Formal extension and complete metrics. After both gates pass, the DPP setting is frozen and its four-round 50K population is generated using the 9140/9141 initial states and branch-noise indices of the earlier formal comparison. The existing QUBO outputs, activations, memberships, and CLIP kernel summaries are reused with their original hashes and protocol identity. This is a post-hoc same-noise operating-point extension. It adds one DPP endpoint to an already observed QUBO population; the earlier two SiT scaling curves remain the independent replications. Each method has 1,000 classes with exactly 50 images per class. The logical budget is 394 denoising NFE/output, with the cached 222-NFE prefix reused during execution.

Table 25: All dynamic formal endpoints on initial/branch namespaces 9140/9141. Every row is a 50K evaluation. QUBO and DPP $\eta = 8$ were generated together; DPP $\eta = 4$ is the two-gate extension.

Selector	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑	CLIP-CMMD↓
Moment-QUBO ($\alpha = 0.56$)	6.51003	8.58418	64.5081	0.84724	0.4046	0.34778
Inception-DPP ($\eta = 4$)	7.70918	8.75581	66.8656	0.84648	0.3929	0.33347
Inception-DPP ($\eta = 8$)	8.72353	9.02330	67.4588	0.86682	0.3595	0.32138

At $\eta = 4$, the formal precision difference is 0.00076 and satisfies the point-matching margin. QUBO improves FID by 1.19914 and recall by 0.01170; DPP has higher IS and lower CLIP-CMMD. The precision interval extends to 0.0058, slightly beyond the positive equivalence boundary. The sFID

interval includes zero (Table 26). These measurements locate an improved Inception FID/recall operating point at similar observed precision while preserving the cross-representation trade-off.

Table 26: Conditional paired 95% intervals on the two dynamic comparisons. Every difference is Moment-QUBO minus DPP at the specified η.

DPP η	Metric	Difference	Conditional 95% interval
4	FID	-1.19914	$[-1.71280, -0.73715]$
4	sFID	-0.17163	$[-0.35420, +0.00581]$
4	Precision	+0.00076	$[-0.00400, +0.00580]$
4	Recall	+0.01170	$[+0.00260, +0.02090]$
4	CLIP-CMMD	+0.01431	$[+0.01067, +0.01776]$
8	FID	-2.21349	$[-2.72486, -1.73528]$
8	sFID	-0.43912	$[-0.64661, -0.23796]$
8	Precision	-0.01958	$[-0.02438, -0.01440]$
8	Recall	+0.04510	$[+0.03610, +0.05400]$
8	CLIP-CMMD	+0.02640	$[+0.02272, +0.02987]$

Evaluation and uncertainty. Calibration streams PNG images into the unchanged ADM feature extractor and tests their precision membership against a fixed 10K real manifold. GPU evaluation disables TF32 and requires exact replay of the original 10K-query membership block; a 100-image PNG extraction smoke check also reproduces all original membership bits. The formal extension uses the same official GPU ADM pipeline, whose P/R membership replay passes exactly. Separate CPU diagnostics retain the small backend-dependent differences.

FID/sFID intervals use paired class influence resampling. Precision resamples generated class-query means while keeping the real manifold fixed; recall resamples the 10K unlabeled real queries with generated manifolds fixed. CLIP-CMMD uses all 50K normalized OpenAI CLIP ViT-L/14@336 image embeddings, Gaussian bandwidth 10, scale 1,000, and the fixed 10K CLIP reference. Its paired class-weight bootstrap uses exact 1,000-by-1,000 class-pair kernel means, rather than subsampling the generated images. All intervals use 2,000 draws and describe composition sensitivity conditional on the archived populations and reference. Independent generation variability and parameter-selection uncertainty are separate sources of uncertainty.

C.15 Class-level trade-offs in the first dynamic comparison

We diagnose the original QUBO–DPP $\eta = 8$ comparison using its fixed 50K caches. QUBO precision is lower in 542 classes, equal in 146, and higher in 312. The ten largest losses account for 6.1% of the total precision-loss mass; 293 classes account for 80%. The difference is spread across many classes. The analysis adds zero diffusion NFE.

Let K^m_{cd} be the mean CLIP Gaussian kernel between the 50 generated images in classes c, d under method m, and let h^m_c be the mean kernel between class c and the real reference. With $C = 1000$ classes and CMMD scale $s = 1000$, define the symmetric additive allocation

$$a_c = s\left[\frac{1}{C}\sum_d (K^Q_{cd} - K^D_{cd}) - 2(h^Q_c - h^D_c)\right], \qquad \Delta\mathrm{CMMD} = \frac{1}{C}\sum_c a_c. \tag{34}$$

Each a_c includes interactions with the other classes and the reference. The allocation reconstructs the formal delta to 7.82×10^{-13}. Of the CMMD difference 0.026398, 0.023858 (90.4%) comes from the generated–generated kernel term, and 0.002540 from the generated–reference term. Its class allocation has Spearman correlation -0.035 with the precision difference. Reweighting away the 100 largest precision-loss classes leaves $\Delta P = -0.00411$ and exact-kernel $\Delta\mathrm{CMMD} = +0.02527$. This post-hoc sensitivity analysis motivates evaluating CLIP separately after precision calibration.

QUBO also has 0.02723 higher final-round normalized lineage entropy and 0.858 more surviving initial roots per class. Lineage records describe ancestry; CLIP kernels describe image-feature distributions. The cached FID influences are zero-centered derivatives with respect to class weights: the ten largest squared influences account for 50.5% of the first-order composition variance. They

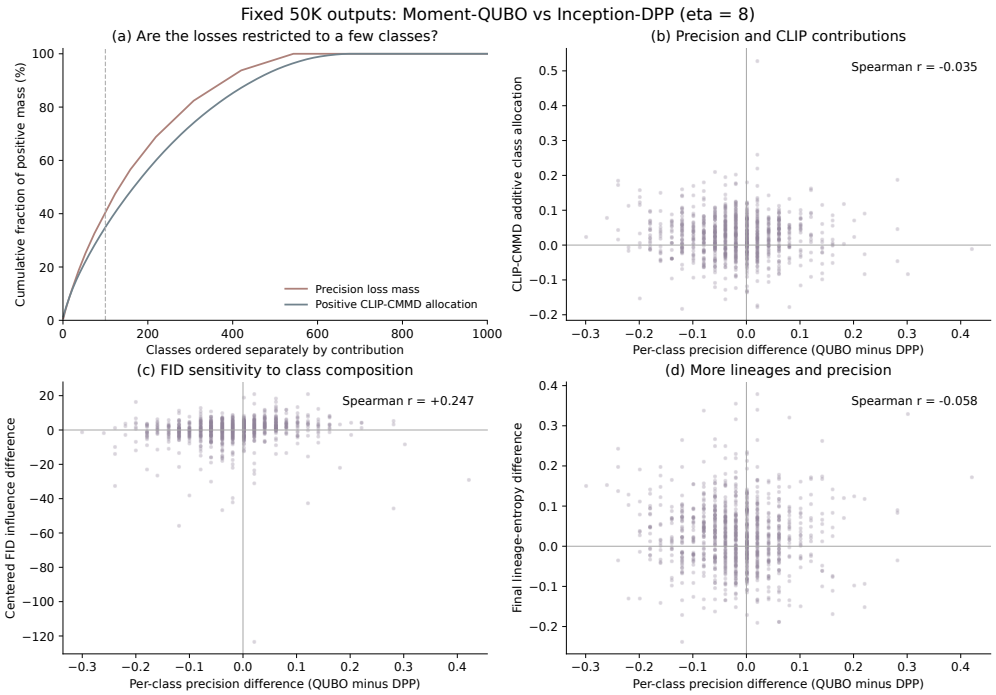

Figure 9: Fixed-output class diagnostics for QUBO versus DPP $\eta = 8$. Loss concentration, exact additive CLIP allocations, centered FID influences, and final-round lineage entropy describe different aspects of the same two archived populations. Each class contains 50 outputs.

measure sensitivity of the full-population FID difference to class composition, rather than class-specific FID values.

C.16 Independent confirmation with frozen dynamic policies

The calibrated same-noise extension in Appendix C.14 determines an operating point. To test its reproducibility on new noise, we freeze Moment-QUBO at $\alpha = 0.56$, $w_1 = w_2 = 1$, and Inception-DPP at $\eta = 4$ before generating either method's confirmation outputs. The registered initial/branch-noise namespaces are 9160/9161; an audit of prior registrations verifies that both are unused. A separate engineering check uses 9162/9163. Each policy runs all four rounds with $K = 50$, $B = 4$, and 394 denoising NFE/output, returning exactly 50 new images for each of the 1,000 ImageNet classes. Methods share the initial states, first-round candidates, and branch-noise indices; subsequent states follow their own retained parents. Every registered endpoint is evaluated regardless of metric ordering, with parameter settings fixed throughout.

FID and recall are the registered primary quality endpoints; sFID, IS, and precision are secondary, and CLIP-CMMD provides a cross-representation evaluation. The precision margin remains 0.005. Both 50K archives pass membership replay, exact cardinality and class-balance checks. The influence calculation reproduces official FID to within 2×10^{-5}.

Table 27: Independent frozen-parameter confirmation: 50K newly generated images per policy, with complete ADM metrics and CLIP-CMMD.

Selector	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑	CLIP-CMMD↓
Moment-QUBO	6.42864	8.43106	64.4062	0.84514	0.4125	0.34791
Inception-DPP ($\eta = 4$)	7.51604	8.53582	66.6519	0.84498	0.3922	0.33535

The QUBO–DPP precision gap is 0.00016, and its conditional interval lies inside $[-0.005, 0.005]$. The 1.0874 FID reduction corresponds to 14.47% relative to DPP; recall increases by 0.0203. DPP has

Table 28: Independent confirmation differences, Moment-QUBO minus Inception-DPP. Intervals condition on the generated archives and reference.

Metric	Difference	Conditional paired 95% interval
FID	-1.08740	$[-1.52152, -0.71307]$
sFID	-0.10476	$[-0.28099, +0.06096]$
Precision	+0.00016	$[-0.00440, +0.00494]$
Recall	+0.02030	$[+0.01050, +0.02980]$
CLIP-CMMD	+0.01256	$[+0.00910, +0.01584]$

higher IS and lower CLIP-CMMD, while the sFID-difference interval crosses zero. This reproduces both the Inception-space benefit and the cross-representation trade-off of the earlier calibrated comparison on a new noise population.

The 2,000 paired resamples use the evaluation procedure in Appendix C.14: class-cluster influences for FID/sFID, generated-class precision means with the real manifold fixed, and the same 10K real query indices for recall with generated manifolds fixed. CLIP-CMMD uses all 50K normalized OpenAI CLIP ViT-L/14@336 embeddings, the fixed 10K reference, Gaussian bandwidth 10 and scale 1,000; paired class weights recompute its full kernel estimator. These intervals measure sensitivity to class/query composition within the frozen artifacts. Precision matching is assessed both by the observed gap and by whether its conditional interval lies inside the margin (Table 28). The independently generated confirmation pair separately tests reproducibility under new generation noise.

C.17 Changing the online verifier to CLIP

To test whether the retention mechanism depends on the DINO quality head, we prospectively specify 128 classes and replace the online verifier with cosine similarity between normalized OpenAI CLIP ViT-L/14@336 image and target-class text embeddings. The single fixed prompt is `a photo of a {class name}.` for each ImageNet label. The Inception self-reference objective, $\alpha = 0.56$, solver, and four-round sampler remain the same. We compare Top-K and Moment-QUBO at $B = 2, 4$, with $K = 50$, initial/branch namespaces 9170/9171, and paired class/noise indices. This yields 6,400 images per configuration and 25,600 in total.

The registered primary diagnostics are final mean pairwise DINO distance and normalized root entropy, defined as in Appendix C.2. DINO embeddings are audit features in this study; CLIP supplies the quality scores used for selection. Verifier score, DINO target logit and Inception distance are secondary diagnostics. Table 29 reports the population means; all differences and the $B = 4$ minus $B = 2$ changes in those differences use 5,000 paired class-bootstrap resamples, seed 20260915176. The reported 95% percentile intervals are descriptive across the specified diagnostics.

Table 29: CLIP online verifier, 128 classes with 50 outputs each per configuration. Distances and entropy describe diversity; the score is target-text CLIP cosine similarity.

B	Selector	CLIP score	DINO dist.	Inception dist.	Root entropy
2	Top-K	0.266116	43.69208	12.93953	0.692534
2	Moment-QUBO	0.265367	45.02391	13.58602	0.698325
4	Top-K	0.284550	39.32721	11.72114	0.535911
4	Moment-QUBO	0.282600	42.03530	12.79369	0.568693

QUBO increases DINO distance at both budgets, by 3.05% at $B = 2$ and 6.89% at $B = 4$. At $B = 4$, root entropy increases by 0.03278 and mean CLIP score decreases by 0.001950, or 0.69%. At $B = 2$, the entropy-difference interval crosses zero. The increases in both the DINO-distance advantage and the entropy advantage from $B = 2$ to $B = 4$ have intervals entirely above zero. This prospective mechanism study extends the verifier comparison within SiT ImageNet search, using 128-class diagnostics as its endpoints. The formal ImageNet FID-50K confirmation above uses the original DINO verifier.

Table 30: CLIP-verifier mechanism differences, QUBO minus Top-K. The $4-2$ rows measure the change in the paired method difference as branching increases.

B	Metric	Difference	Paired 95% interval
2	DINO distance	+1.331836	$[+0.994673, +1.691360]$
2	Root entropy	+0.005792	$[-0.001391, +0.012818]$
2	Inception distance	+0.646484	$[+0.513696, +0.777166]$
2	Verifier score	-0.000749	$[-0.001463, -0.000027]$
4	DINO distance	+2.708091	$[+2.331870, +3.110865]$
4	Root entropy	+0.032782	$[+0.021373, +0.044740]$
4	Inception distance	+1.072543	$[+0.913937, +1.235865]$
4	Verifier score	-0.001950	$[-0.002527, -0.001385]$
$4-2$	DINO distance	+1.376255	$[+0.951064, +1.804767]$
$4-2$	Root entropy	+0.026991	$[+0.013241, +0.040379]$
$4-2$	Inception distance	+0.426059	$[+0.240242, +0.621209]$
$4-2$	Verifier score	-0.001201	$[-0.002080, -0.000339]$

C.18 Score-only stochastic retention on reserved classes

A softer score-only rule can preserve alternatives even without pairwise population terms. We implement sampling without replacement through

$$S_\tau = \mathrm{TopK}_i\left(\widetilde{q}_i/\tau + g_i\right), \qquad g_i \overset{\text{iid}}{\sim} \mathrm{Gumbel}(0,1), \quad \tau > 0, \tag{35}$$

where $\widetilde{q}$ is the same median/MAD-standardized DINO quality score used by the main objective. Gumbel perturbations are freshly seeded for each class, budget and round. Selection retains exactly 50 distinct candidates and their latent states, which become the next round's parents. This is a score-only retention control inside the existing four-round path search. We keep $B = 4$, QUBO $\alpha = 0.56$, initial/branch namespaces 9172/9173 and selector seed 20260915172 fixed.

Screening and amendment chronology. The original plan reserves 64 screening classes and 128 disjoint held-out classes before generation. Screening evaluates $\tau \in \{0.25, 0.5, 1\}$ and QUBO, with 3,200 images per configuration. It selects the closest precision point to QUBO and requires an absolute gap at most 0.005 to continue. The closest temperature, $\tau = 0.5$, has a gap of 0.0090625, so the registered screening gate fails and the original queue stops. Table 31 retains all three settings.

After inspecting this failed screen, we register an amendment to evaluate the same $\tau = 0.5$ and QUBO on the originally reserved 128 classes. Both policies and the class set are frozen before any continuation outputs are generated. The amendment requests completion regardless of ordering and allows no retuning on reserved classes. Thus this study is an amended follow-up on previously unobserved classes, with the original screening decision preserved. Each method produces 6,400 images, totaling 12,800.

Table 31: Original 64-class temperature screen. Differences are QUBO minus sampling. Every precision point gap exceeds the registered 0.005 margin.

τ	QUBO precision	Sampling precision	Difference	Paired 95% interval
0.25	0.870000	0.896563	-0.026562	$[-0.044383, -0.009680]$
0.5	0.870000	0.879062	-0.009062	$[-0.028125, +0.009695]$
1	0.870000	0.836875	+0.033125	$[+0.014687, +0.051258]$

Table 32: Fixed-policy continuation on the original 128 reserved classes, 6,400 images per method. The quality score is the DINO target-class logit.

Selector	Precision	Verifier score	DINO dist.	Inception dist.	Root entropy
Moment-QUBO	0.838906	18.367878	36.86858	11.84188	0.573104
Sampling ($\tau = 0.5$)	0.842344	17.707197	37.65558	11.34405	0.585328

Table 33: Reserved-class differences, QUBO minus sampling at $\tau = 0.5$. Precision conditions on the fixed real manifold; mechanism intervals use the same paired class resamples.

Metric	Difference	Paired 95% interval
Precision	-0.003437	$[-0.017344, +0.009844]$
DINO distance	-0.787000	$[-1.192433, -0.389832]$
Root entropy	-0.012224	$[-0.026601, +0.001612]$
Inception distance	+0.497829	$[+0.353017, +0.641492]$
Verifier score	+0.660680	$[+0.583466, +0.741450]$

The continuation's precision difference is -0.0034375, so its absolute value is below 0.005. Its interval $[-0.017344, 0.009844]$ extends beyond both boundaries of the matching region. Sampling gives higher DINO distance by 0.7870; QUBO gives higher DINO verifier score by 0.6607 and Inception distance by 0.4978. Each of these three difference intervals excludes zero. Sampling's root-entropy mean is higher by 0.01222, with an interval crossing zero. These results place stochastic retention on a different quality–diversity operating point from moment matching, and show that softened pointwise ranking also preserves diverse paths. The evidence consists of official ADM precision and paired mechanism diagnostics on 128 classes. Complete generated rows, compact numerical evidence and source hashes accompany the manuscript as `strengthening_rows.tex`, `strengthening_evidence.json` and `strengthening_manifest.json` in the `generated/` directory.

C.19 Fixed-temperature extension to all 1,000 classes

We evaluate the fixed $\tau = 0.5$ retention rule in Eq. 35 on 50,000 images, with exactly 50 outputs for each ImageNet class. The comparison reuses the initial/branch-noise namespaces 9160/9161 and the frozen Moment-QUBO archive from Appendix C.16. Because the QUBO results were already available when this comparison was registered, it is a post-hoc same-noise comparison. The separate 9172/9173 studies determine τ before these 50K outputs are generated. We retain $K = 50$, $B = 4$, four pruning rounds, $\alpha = 0.56$, and selector seed 20260915172, and evaluate all metrics regardless of precision matching or metric ordering.

Table 34: Full ImageNet-50K comparison against fixed-temperature stochastic retention. Each method has 50,000 images; the QUBO archive and its original evaluation are reused. Both policies have the same logical denoising budget of 394 NFE per output.

Selector	FID ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑	CLIP-CMMD ↓
Moment-QUBO	6.42864	8.43106	64.4062	0.84514	0.4125	0.34791
Sampling ($\tau = 0.5$)	7.07040	8.53271	64.4410	0.85840	0.4088	0.35891

Table 35: 50K endpoint differences, Moment-QUBO minus sampling at $\tau = 0.5$. Intervals condition on the frozen archives and reference. IS is reported as an official point estimate in Table 34.

Metric	Difference	Conditional paired 95% interval
FID	-0.641763	$[-0.938906, -0.338072]$
sFID	-0.101646	$[-0.252708, +0.043370]$
Precision	-0.013260	$[-0.018101, -0.008540]$
Recall	+0.003700	$[-0.006000, +0.012802]$
CLIP-CMMD	-0.011000	$[-0.014644, -0.007465]$

Moment-QUBO lowers FID from 7.07040 to 6.42864, a 9.08% reduction, and lowers CLIP-CMMD from 0.35891 to 0.34791. Both difference intervals exclude zero. Its precision is lower by 0.01326; the conditional interval for the QUBO-minus-sampling difference is $[-0.018101, -0.008540]$. The point difference and the entire interval lie below the registered ± 0.005 matching region. The sFID and recall difference intervals cross zero. These 50K endpoints place moment matching and softened

Table 36: Prospectively specified quadratic-structure ablation on an independent fresh-noise 200K-candidate pool. All methods select 50K images from the same decoded candidates and add no diffusion NFE. Full Moment-QUBO denotes the global real-reference moment objective.

Selector	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
Inception-DPP	7.9562	8.6699	66.8233	0.8604	**0.3825**
Full Moment-QUBO	**7.9129**	**8.6579**	66.8115	0.8601	0.3822
Linear-only	7.9767	8.6669	66.8339	**0.8609**	0.3820
Shuffled quadratic	7.9711	8.6745	**66.8442**	0.8608	0.3820

Table 37: Two independent DiT-XL/2 ImageNet-256 50K RBF-control replications. Candidates are shared across selectors within each pool; each selector uses 110 network-equivalent NFE/output.

Pool	Selector	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
1	Inception-DPP	4.2337	6.9964	**61.8877**	**0.8476**	0.4919
	RBF-MMD-QUBO	**4.0217**	**6.4845**	61.7507	0.8465	**0.4940**
2	Inception-DPP	4.2387	6.9245	**61.5657**	**0.8462**	0.4953
	RBF-MMD-QUBO	**4.0128**	**6.4154**	61.5205	0.8435	**0.5025**

pointwise ranking at distinct operating points: QUBO favors distributional agreement in the two evaluated feature spaces, while sampling favors ADM precision.

The confidence procedure is identical to the frozen 50K confirmation: 2,000 paired resamples, class-cluster influence intervals for FID/sFID, generated-class resampling for precision with the real manifold fixed, and shared real-query resampling for recall with generated manifolds fixed. CLIP-CMMD uses exact class-pair kernel summaries of the complete 50K estimator against the fixed 10K reference. These intervals describe class/query-composition sensitivity conditional on the completed archives. The comparison retains the temperature-screen failure and 128-class continuation as preceding studies.

The extension reuses the frozen 222-NFE prefix and first-round feature cache. Before generation, an excluded engineering replay verifies bitwise equality of selected indices, final pixels and features, and initial/branch lineages between cached and uncached executions. Each new output consequently requires 168 additional denoiser evaluations; both policies retain their full logical 394-NFE accounting. This reuse reduces the computation required to add the experimental endpoint. The archived QUBO evaluation is verified by source hashes, and both ADM membership replays reproduce precision and recall exactly. The manuscript includes the generated rows, compact evidence and source hashes as `temperature_formal_rows.tex`, `temperature_formal_evidence.json` and `temperature_formal_manifest.json` in `generated/`.

C.20 Quadratic-structure controls

Before generating the fresh candidate pool, the protocol fixed objective weights, per-class Inception-DPP initialization, cardinality, 0.995 quality floor, swap budget, shuffle namespace, and image gates. Every method selects 50K from the same 200K decoded images. Linear-only deletes off-diagonal terms; shuffled quadratic preserves the linear terms and pair-coefficient multiset while permuting candidate identities within each class. The selection comparisons add zero diffusion NFE. Here "Full Moment-QUBO" is the global real-reference objective, distinct from candidate-self-reference dynamic retention. Linear-only retains its binary linear coefficients; it is not the mean-only moment ablation.

C.21 Calibrated DiT RBF-MMD-QUBO selection

C.22 DiT RBF objective and solver decomposition

Table 38: Same-pool DiT comparison separating the RBF-MMD population objective from one-swap refinement.

Pool	Selector	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
1	Inception-DPP	4.233657	6.996400	61.887672	0.847600	0.491900
	RBF-MMD-greedy	**4.017949**	**6.477331**	61.771584	0.846480	**0.494300**
	RBF-MMD-one-swap	4.021668	6.484506	61.750710	0.846520	0.494000
2	Inception-DPP	4.238717	6.924527	**61.565674**	**0.846220**	0.495300
	RBF-MMD-greedy	4.018334	**6.414566**	61.513390	0.843620	**0.502800**
	RBF-MMD-one-swap	**4.012762**	6.415387	61.520466	0.843460	0.502500

On pool 2, greedy and one-swap choose different sets in 143 of 1,000 classes and differ in only 160 of 50,000 selected slots. One or more improving swaps are accepted in 148 classes. Across the 917 classes that choose the same α, one-swap strictly reduces the frozen QUBO energy in 88 classes, with mean energy change -6.264×10^{-6}. The small selection difference explains why downstream metrics are nearly identical.

D Additional Retention Controls

The following experiments distinguish candidate-moment matching from generic pairwise diversity, separate first- and second-moment terms, and test repeated branching with a second backbone. The SiT controls evaluate new retention policies on the existing noise bank used by the independent scaling curve. The DiT pilot uses a separately registered generation protocol. Independent full-class confirmations of the frozen SiT and DiT policies follow in Appendix E.

D.1 Shared protocol and calibration

All SiT controls use the same pretrained SiT-XL/2, DINO verifier, four retention rounds, $K = 50$, $B = 4$, and 394 denoising NFE/output. Initial states and branch-noise indices are shared across policies. Identical parents at the same round reuse identical candidates; once parents diverge, their candidates are generated separately. Sixteen classes are used for calibration and 64 disjoint classes for the initial evaluation. These classes are selected by a frozen hash order.

We adapt the quality–diversity objective of Group Inference (Parmar et al., 2026) to this shared branching framework:

$$E_{\mathrm{GI}}(x) = -\frac{1}{K}\sum_i \widetilde{q}_i x_i - \frac{2\lambda}{K(K-1)}\sum_{i<j}\big(1-\cos(f_i,f_j)\big)x_i x_j, \qquad \mathbf{1}^\top x = K. \tag{36}$$

Here $\widetilde{q}_i$ is the same robustly standardized DINO target-class logit used by Moment-QUBO. This GI-style control isolates a quality plus pairwise-repulsion objective under our sampler, features, and solver; the original Group Inference system uses its own generation and optimization pipeline. Both quadratic objectives use constructive greedy followed by best one-out/one-in exchanges, with the same 100-scan cap and 10^{-12} improvement tolerance.

For each candidate setting, we compute the mean standardized verifier score of the retained set in each calibration class, then average these scores over the 16 classes. We select the setting with the smallest absolute difference from Moment-QUBO's average on the same first-round candidates. Calibration uses neither final-image FID nor ADM precision. Each family has eight candidate settings: $\lambda \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16\}$ for GI-style; $\alpha \in \{0.07, 0.14, 0.28, 0.56, 1.12, 2.24, 4.48, 8.96\}$ for moment variants; and quality weights $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.75, 0.9\}$ for MMR. Ties follow the registered grid order. Selected settings are $\lambda = 2$ for Inception GI-style, $\lambda = 1$ for DINO GI-style, MMR quality weight 0.4, and $\alpha = 0.56$ for mean-only and DINO-moment matching. Second-only matching selects $\alpha = 0.07$, the grid boundary, with a remaining standardized-score difference

of 0.02151. The fixed-α moment ablations are reported separately from this calibrated variant. Calibration controls the first-round verifier operating point; final ADM precision remains an evaluated outcome.

D.2 Full ImageNet comparison with pairwise diversity

After the 64-class pilot, the Inception GI-style and mean-only policies are frozen and extended to all 1,000 classes, with exactly 50 images per class. Their 6,400 pilot images are reused and 93,600 images are added. Existing Top-K and Moment-QUBO outputs from the same noise bank supply the paired controls, and all four 50K archives are evaluated consistently with the official ADM pipeline. This all-class extension includes the 16 calibration and 64 pilot classes. Its noise population is the existing SiT bank, rather than an additional independent replication.

Table 39: All-class SiT retention controls, $B = 4$: 50K outputs per policy on the same noise bank. New GI-style and mean-only policies are frozen after the pilot. Entries are full ADM point estimates.

Retention policy	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
Top-K	9.525122	9.170535	67.3279	0.88326	0.3290
Moment-QUBO, $w_1 = w_2 = 1$	6.455375	8.528964	64.2655	0.84356	0.4124
Inception GI-style, $\lambda = 2$	7.374812	8.980665	65.8977	0.84068	0.4078
Mean-only, $w_1 = 1, w_2 = 0$	6.459215	8.627116	64.4788	0.84270	0.4024

Moment-QUBO improves FID over GI-style by 0.91944 (12.47%), with precision higher by 0.00288 and recall by 0.0046. GI-style has higher IS. At these frozen operating points, the conditional FID and sFID intervals favor Moment-QUBO, while precision and recall intervals cross zero (Table 40). This supports an Inception-space distributional improvement over the calibrated pairwise-diversity control. Cross-representation comparisons below further delimit that advantage.

Paired uncertainty on the frozen 50K populations. After the point estimates were known, we registered a statistical analysis that reuses the three generated archives, re-extracts their official ADM features, and verifies agreement with the frozen metrics. Each interval uses 20,000 paired bootstrap draws. FID and sFID use a first-order class-influence approximation, anchored at the official difference and resampling the same 1,000 generated classes in both methods. Precision resamples the paired class means of generated-image membership indicators, with the real manifold fixed. Recall resamples paired membership indicators for the same 10,000 real queries, with both generated manifolds fixed; the reference archive has no class labels for class-level resampling. These descriptive 95% intervals quantify composition sensitivity conditional on the observed outputs and real reference. They exclude regeneration, tuning uncertainty, and manifold refitting, and are unadjusted for multiple comparisons. An interval crossing zero leaves the direction unresolved; it does not establish equivalence. The analysis adds no diffusion evaluations and retains the same-noise extension's scope, including its calibration and pilot classes.

Table 40: Paired conditional uncertainty for the frozen 50K retention controls. Each difference is Moment-QUBO minus the named baseline. FID/sFID are lower-is-better; precision/recall are higher-is-better. The intervals use the metric-specific resampling units described above.

Baseline	Metric	Difference	Conditional 95% interval
Inception GI-style	FID	-0.9194372	$[-1.2360463, -0.6256705]$
Inception GI-style	Precision	$+0.0028800$	$[-0.0015000, +0.0073000]$
Inception GI-style	Recall	$+0.0046000$	$[-0.0046000, +0.0138000]$
Inception GI-style	sFID	-0.4517004	$[-0.6374221, -0.2673461]$
Mean-only	FID	-0.0038396	$[-0.1591154, +0.1725217]$
Mean-only	Precision	$+0.0008600$	$[-0.0026400, +0.0043800]$
Mean-only	Recall	$+0.0100000$	$[+0.0018000, +0.0182000]$
Mean-only	sFID	-0.0981514	$[-0.2008055, +0.0043361]$

To separate diagnostics from the calibration and pilot cohorts, we also report paired contrasts on the remaining 920 classes (Table 41). Each policy's moment discrepancy uses its own current candidate

population as reference; after parent selection diverges, these references differ. The intervals use 2,000 paired class-bootstrap resamples conditional on this noise bank. They quantify class-level mechanism diagnostics, separately from the full 50K distributional metrics.

Table 41: GI-style minus Moment-QUBO on 920 classes outside calibration and the initial pilot. Root entropy is normalized by $\log K$.

Final-round diagnostic	Difference	Paired 95% interval
Mean DINO verifier	$+0.13300$	$[0.11309, 0.15402]$
Root entropy	-0.01592	$[-0.02045, -0.01170]$
Pairwise DINO distance	-0.28515	$[-0.40767, -0.16851]$
Selected-to-own-candidate moments	$+0.01411$	$[0.01275, 0.01555]$

D.3 Moment terms and classical diversity controls

The moment ablation compares $(w_1, w_2) = (1, 0), (0, 1), (1, 1)$ at the same $\alpha = 0.56$. Statistical moment order differs from polynomial degree in the binary variables: even first-moment matching remains quadratic, because

$$\left\| \frac{1}{K} \sum_i x_i f_i - \frac{1}{M} \sum_i f_i \right\|_2^2 = \frac{1}{K^2} \sum_{i,j} x_i x_j f_i^\top f_j - \frac{2}{KM} \sum_{i,j} x_i f_i^\top f_j + C. \tag{37}$$

Reference centering and normalization rescale this discrepancy as in Equation 4. Thus this ablation separates statistical targets, whereas the linear-only and shuffled-interaction controls elsewhere separate binary-objective terms.

Table 42 reports all 12 pilot configurations. Mean-only closely follows the full moment objective in FID; second-only is substantially weaker at the fixed quality weight, and remains weaker after the first-round calibration described above. The 50K extension gives nearly identical mean-only and full-objective FID point estimates: 6.459215 versus 6.455375, a difference of 0.00384. Adding second moments increases recall by 0.0100, with conditional interval $[0.0018, 0.0182]$; FID, sFID, and precision intervals cross zero (Table 40). Mean matching therefore captures nearly all the observed FID benefit in this control. The fresh-noise confirmation in Appendix E.1 repeats the recall gain, while its FID interval also spans zero. The evidence favors population-statistic matching over Top-K while assigning a more limited role to the second-moment term.

MMR (Carbonell & Goldstein, 1998) sequentially balances percentile verifier quality against maximum cosine similarity to the selected set. The k-medoids control (Kaufman & Rousseeuw, 1990) uses PAM build-and-swap with at most 20 passes to minimize nearest-medoid squared Inception distance. It is a quality-free geometric reference: its high recall is accompanied by markedly lower precision. All policies operate inside the same repeated branching loop, including the final round.

Table 42: Complete 64-class SiT pilot, 3,200 outputs per policy. FID and recall compare the class-restricted generated cohort with the full official ImageNet reference; their magnitudes are not directly comparable to the corresponding metrics from the all-class 50K evaluations. DINO-KID and CLIP-CMMD are frozen-output audits on this same cohort. FID, DINO-KID, and CLIP-CMMD are lower-is-better.

Policy	FID	Prec.	Rec.	Root entropy	DINO dist.	DINO-KID	CLIP-CMMD
Top-K	92.03114	0.90813	0.0238	0.50127	34.4821	0.278153	0.347822
Moment-QUBO $(1, 1)$	77.50383	0.87094	0.0447	0.53424	36.9438	0.265045	0.366890
Mean-only $(1, 0)$	77.75536	0.87469	0.0468	0.52620	36.8754	0.264537	0.368198
Second-only $(0, 1)$	90.33360	0.90063	0.0265	0.51147	34.5832	0.275059	0.345280
Inception-DPP, greedy	82.61239	0.86125	0.0346	0.53370	36.1146	0.270470	0.363818
Inception-DPP, greedy + swap	82.61134	0.86125	0.0346	0.53370	36.1187	0.270470	0.363832
Inception k-medoids	82.77796	0.38688	0.3941	0.63427	53.9829	1.114884	1.363538
Inception GI-style, $\lambda = 2$	80.85096	0.87156	0.0416	0.52227	36.5487	0.262310	0.350264
DINO GI-style, $\lambda = 1$	86.25676	0.89531	0.0425	0.51686	37.1108	0.256351	0.341056
Inception MMR, quality weight 0.4	80.18176	0.86563	0.0494	0.53735	36.4855	0.267332	0.358065
Second-only, calibrated $\alpha = 0.07$	88.88665	0.90469	0.0305	0.53369	35.2514	0.271149	0.346804
DINO moments, calibrated $\alpha = 0.56$	79.05430	0.86875	0.0427	0.55707	39.3677	0.254216	0.377497

DPP refinement fairness. The DPP refinement control starts from the identical greedy MAP solution on the same quality-weighted Inception RBF kernel, with $\eta = 4$ and diagonal jitter 10^{-8}. At each scan it evaluates every one-out/one-in exchange by its log-determinant gain and accepts the best gain exceeding 10^{-12}, up to 100 scans. This matches the exchange neighborhood and scan cap of Moment-QUBO, with each method optimizing its own objective. It does not match elapsed solver time. Parent slots retain their greedy order, and an accepted exchange replaces only the removed slot, preserving subsequent noise assignments when no exchange is accepted. Across 256 class-rounds, three rounds accept exchanges, with four accepted exchanges in total. Pilot FID changes by -0.00105; precision and recall are unchanged at the reported precision. The added refinement therefore has little effect on this DPP operating point.

D.4 Cross-representation audit of the controlled policies

The last two columns of Table 42 evaluate every frozen pilot archive against the same 10K ImageNet reference. DINO-KID uses the all-sample unbiased cubic-kernel estimator $k(u, v) = (u^\top v/1024 + 1)^3$. CLIP-CMMD uses normalized OpenAI CLIP ViT-L/14@336 embeddings, Gaussian bandwidth 10, and scale 1,000. The audit retains all 12 settings; it follows generation rather than defining a new prospective evaluation cohort. Inception and DINO moment selectors provide two online representations evaluated in three feature spaces. DINO additionally supplies the shared quality verifier; online CLIP-moment selection is outside this comparison.

For uncertainty, let $\widehat{\Delta}_{(-c)}$ be a metric difference after deleting generated class c from both policies, keeping the real reference fixed. With $C = 64$, the paired jackknife standard error is

$$\widehat{\mathrm{SE}}_{\mathrm{jack}} = \left[\frac{C-1}{C} \sum_{c} \left(\widehat{\Delta}_{(-c)} - \overline{\Delta}_{(-\cdot)} \right)^2 \right]^{1/2}. \tag{38}$$

We report approximate intervals centered on the observed full-cohort difference, $\widehat{\Delta} \pm t_{63,0.975} \widehat{\mathrm{SE}}_{\mathrm{jack}}$. An initial calculation retained raw class-resampling percentiles, which can shift with class mixture for quadratic statistics. The reviewed jackknife intervals replace those percentiles for uncertainty reporting; point estimates are unchanged. These intervals describe cohort sensitivity conditional on the frozen outputs and real reference.

Table 43: Cross-representation differences, Moment-QUBO minus the comparison policy. Approximate paired 95% class-jackknife intervals condition on the fixed 64-class outputs and 10K reference.

Comparison policy	Metric	Difference	Approximate 95% interval
Top-K	DINO-KID	-0.013108	$[-0.020896, -0.005321]$
Top-K	CLIP-CMMD	$+0.019067$	$[0.001488, 0.036646]$
Inception GI-style	DINO-KID	$+0.002735$	$[-0.003366, 0.008836]$
Inception GI-style	CLIP-CMMD	$+0.016626$	$[0.002211, 0.031042]$
Mean-only	DINO-KID	$+0.000508$	$[-0.002849, 0.003865]$
Mean-only	CLIP-CMMD	-0.001309	$[-0.011047, 0.008430]$

Inception-moment retention improves DINO-KID over Top-K in this cohort, while CLIP-CMMD favors Top-K and GI-style. DINO-moment retention lowers DINO-KID further but has a worse CLIP-CMMD point estimate than Inception moments. Mean-only and full-moment cross-representation differences have intervals spanning zero. Together with the independent matched-precision DPP comparison, these controls establish a representation-dependent trade-off rather than a single ordering across feature spaces.

D.5 Repeated branching with DiT

This pilot implements repeated branching with the official DiT-XL/2 ImageNet-256 checkpoint, distinct from the earlier single-prune and fully denoised pool experiments. We use 25-step deterministic DDIM, classifier-free guidance scale 1.5, the MSE-trained latent VAE, and $K = 50$. Thirty-two classes are fixed by hash order before generation, giving 1,600 outputs per policy. An additional disjoint class checks the implementation. Splitting the base trajectory into segments reproduces the unmodified DDIM25 output with a maximum absolute latent difference of zero.

After ten shared DDIM steps, each retained state is independently re-noised into B children using the VP transition

$$z_s = \sqrt{\frac{\bar{\alpha}_s}{\bar{\alpha}_t}} z_t + \sqrt{1 - \frac{\bar{\alpha}_s}{\bar{\alpha}_t}} \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \quad s > t. \tag{39}$$

Each child then follows deterministic DDIM to the next retention point. On the respaced schedule, the four (t, s, end) triples are $(14, 16, 10)$, $(10, 12, 6)$, $(6, 8, 2)$, and $(2, 4, -1)$, where -1 denotes the clean endpoint. Intermediate clean predictions provide selection features; the retained states become the next round's parents. Branch noise is indexed by class, round, parent slot, and branch, with nested branch indices across budgets. The four branch segments use 23 DDIM steps in total. Counting both classifier-free guidance network evaluations gives $2(10 + 23B)$ NFE/output.

Moment-QUBO uses the SiT settings $\alpha = 0.56$, $w_1 = w_2 = 1$; Inception-DPP uses $\eta = 4$. These values are transferred directly, with no DiT parameter fitting or ADM precision calibration. The $B = 1$ identity retention control still re-noises states, and is therefore distinct from the plain DDIM25 sampler.

Table 44: DiT repeated-branching pilot: 32 fixed classes, 1,600 images per policy. FID and recall use the full official ImageNet reference, so absolute values apply to this class-restricted cohort.

Policy	NFE	FID↓	Prec.↑	Rec.↑	Root entropy	DINO dist.
DDIM25 base	50	80.91046	0.83063	0.0984	1.00000	45.0708
$B = 1$, identity	66	83.88263	0.85375	0.0807	1.00000	44.3482
$B = 2$, Top-K	112	126.85070	0.96375	0.0089	0.67341	33.9322
$B = 2$, Moment-QUBO	112	125.38193	0.96250	0.0109	0.68513	34.2869
$B = 2$, Inception-DPP	112	123.85884	0.95563	0.0114	0.66797	34.4307
$B = 4$, Top-K	204	137.79856	0.95438	0.0058	0.49613	31.2862
$B = 4$, Moment-QUBO	204	132.84614	0.95875	0.0082	0.52593	31.8923
$B = 4$, Inception-DPP	204	131.63236	0.95063	0.0077	0.52434	32.2611

The Moment-QUBO FID advantage over Top-K increases from 1.46877 at $B = 2$ to 4.95242 at $B = 4$. All three selection policies deteriorate with larger branching, DPP has the lowest FID among them, and the base sampler retains substantially better FID and recall than the searched outputs. Thus the pilot supports relative mitigation of the retention failure in a second repeated-branching system. Its sample size and restricted class coverage limit comparison with the main 50K study. Paired class-bootstrap intervals also show uncertainty in the mechanism diagnostics: the QUBO-minus-Top-K root-entropy differences are 0.01172 $[-0.00076, 0.02370]$ at $B = 2$ and 0.02981 $[-0.00147, 0.06012]$ at $B = 4$; the corresponding DINO-distance intervals also span zero. Shared-prefix execution times describe group throughput, rather than isolated per-policy latency. The independently generated all-class 50K results in Appendix E.2 provide the full-population test of these frozen policies.

Numerical evaluation amendment. With 1,600 samples and 2,048-dimensional Inception features, generated covariance matrices are singular. The initial nonsymmetric covariance-product square root produced large imaginary components and failed evaluation. A recorded amendment evaluates the same Gaussian Fréchet formula through the symmetric positive-semidefinite matrix $A = \Sigma_r^{1/2} \Sigma_g \Sigma_r^{1/2}$:

$$\text{FID} = \|\mu_g - \mu_r\|_2^2 + \text{Tr}(\Sigma_g + \Sigma_r) - 2 \sum_j \sqrt{\lambda_j(A)}. \tag{40}$$

Float64 symmetric eigendecomposition clips only negative eigenvalues within $10^{-9} \max(1, \max_j |\lambda_j|)$ of zero and rejects more negative values; no covariance ridge is added. Official ADM feature extraction, reference statistics, IS, and precision/recall functions are retained. Analytic singular/full-rank checks pass. Evaluator validation on cached features from a formal 50K run reproduces official FID within 7.96×10^{-13} and sFID within 1.09×10^{-7}. The amendment applies to this pilot; the formal SiT 50K tables use the original evaluation procedure.

Artifacts. The accompanying result bundles `reviewer_dynamic_controls_20260920`, `reviewer_full_confirmation_20260920`, `reviewer_crossfeature_20260920`,

and `dit_repeated_branching_20260920` contain the frozen protocols, calibration choices, full metric records, class diagnostics, and sample hashes. Cross-representation intervals use `uncertainty_review/summary.json`; the DiT numerical amendment and full-rank regression are recorded under `recovery_20260921`.

E Independent Full-Class Retention Confirmations

Two prospective studies test the frozen retention policies on newly generated noise populations. The SiT study compares moment matching with the calibrated pairwise-diversity control and with mean-only matching; the DiT study tests repeated branching on a second backbone. Every configuration produces 50 images in each of 1,000 ImageNet classes. Initial and branch noises are independent of the preceding studies and shared across policies within each new study. At a fixed branching budget, later candidate populations diverge as their retained parents change.

E.1 Fresh-noise SiT retention controls

We freeze the three settings from Appendix D: Moment-QUBO with $\alpha = 0.56$, $w_1 = w_2 = 1$; Inception GI-style with $\lambda = 2$; and mean-only matching with $\alpha = 0.56$, $w_1 = 1, w_2 = 0$. All use the same four-round SiT search, $K = 50$, $B = 4$, verifier, representation, and greedy-plus-swap solver, at 394 NFE/output. The new initial/branch namespaces are 202609220/202609221, and all 150K outputs are newly generated. Hyperparameters and metric comparisons are registered before generation. Table 45 contains all three formal 50K evaluations.

Table 45: Independent fresh-noise SiT confirmation, $B = 4$. Each row is an official class-balanced ImageNet-256 ADM-50K evaluation under a frozen retention policy.

Retention policy	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
Moment-QUBO, $w_1 = w_2 = 1$	6.568306	8.562873	64.4574	0.84610	0.4084
Inception GI-style, $\lambda = 2$	7.496022	8.955610	65.7778	0.84626	0.3978
Mean-only, $w_1 = 1, w_2 = 0$	6.653289	8.650646	64.4893	0.84560	0.3953

Moment-QUBO improves FID over GI-style by 0.92772, with precision difference -0.00016 and recall gain 0.0106. The conditional intervals favor Moment-QUBO for FID, sFID, and recall, while the precision interval spans zero (Table 46). Thus the Inception-space advantage over calibrated pairwise repulsion repeats on new noise. Against mean-only matching, the FID difference is -0.08498 with an interval spanning zero; the recall gain is 0.0131 with interval $[0.0048, 0.0213]$. Together with the preceding cohort, this identifies mean matching as the main source of the FID benefit and a smaller, repeatable recall contribution from adding second moments.

Table 46: Independent SiT confirmation: Moment-QUBO minus each frozen baseline. Two-sided 95% intervals describe conditional class/query composition sensitivity.

Baseline	Metric	Difference	Conditional 95% interval
Inception GI-style	FID	-0.92772	$[-1.21729, -0.65582]$
Inception GI-style	sFID	-0.39274	$[-0.57116, -0.21180]$
Inception GI-style	Precision	-0.00016	$[-0.00454, 0.00430]$
Inception GI-style	Recall	$+0.01060$	$[0.00140, 0.01960]$
Mean-only	FID	-0.08498	$[-0.25489, 0.12076]$
Mean-only	sFID	-0.08777	$[-0.19372, 0.01681]$
Mean-only	Precision	$+0.00050$	$[-0.00294, 0.00402]$
Mean-only	Recall	$+0.01310$	$[0.00480, 0.02130]$

E.2 Full-class repeated branching with DiT

We transfer the eight pilot configurations in Appendix D.5 unchanged to a fresh, complete ImageNet population: DDIM25 base, $B = 1$ identity retention, and Top-K, Moment-QUBO,

and Inception-DPP at both $B = 2$ and $B = 4$. The sampler, four re-noising/retention nodes, $K = 50$, guidance 1.5, QUBO $\alpha = 0.56$, and DPP $\eta = 4$ remain fixed. The namespace `dit-repeated-confirmation-20260922-v1` separates formal initial and branch seeds from the earlier pilot and implementation checks. Within this study, $B = 2$ branch indices are nested inside $B = 4$. The eight configurations generate 400K outputs in total.

Table 47: Independent DiT repeated-branching confirmation: 50K outputs per row. NFE counts both classifier-free guidance evaluations. The $B = 1$ control includes re-noising; DDIM25 base uses plain sampling. Root entropy is normalized by $\log 50$.

Policy	NFE	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑	Root entropy
DDIM25 base	50	3.035812	4.786885	57.6474	0.78976	0.5806	1.00000
$B = 1$, identity	66	2.950660	5.183428	59.5769	0.80404	0.5620	1.00000
$B = 2$, Top-K	112	12.940944	9.979923	71.3218	0.93480	0.2174	0.65804
$B = 2$, Moment-QUBO	112	12.161709	9.380374	71.2006	0.93314	0.2369	0.66801
$B = 2$, Inception-DPP	112	11.873350	9.616727	71.5105	0.92784	0.2529	0.66629
$B = 4$, Top-K	204	16.916425	12.726302	73.4829	0.93746	0.1231	0.47654
$B = 4$, Moment-QUBO	204	15.356187	11.646177	73.2922	0.93866	0.1505	0.51347
$B = 4$, Inception-DPP	204	14.866351	12.043874	73.4751	0.92022	0.1697	0.49845

Moment-QUBO's FID advantage over Top-K grows from 0.77924 at $B = 2$ to 1.56024 at $B = 4$, while its recall gain grows from 0.0195 to 0.0274. All four conditional intervals exclude zero (Table 48). Its final root-entropy gains are 0.00997 $[0.00756, 0.01238]$ and 0.03693 $[0.03208, 0.04170]$, respectively; paired class-bootstrap intervals also favor its DINO pairwise distance. This full-class experiment reproduces the relative retention benefit in a second repeated-branching system.

Both population selectors still deteriorate from $B = 2$ to $B = 4$, and plain DDIM25 retains better FID and recall than all $B \geq 2$ search configurations. Inception-DPP has lower FID point estimates and higher recall than Moment-QUBO, while Moment-QUBO has higher precision and lower sFID. Their FID intervals span zero, whereas the precision/recall intervals reflect distinct operating points. These fixed transferred policies test cross-backbone retention behavior at their observed precision levels.

Table 48: Independent DiT confirmation: Moment-QUBO minus the named policy at the same branching budget. All intervals are conditional 95% intervals on paired class/query composition.

Comparison	Metric	Difference	Conditional 95% interval
$B = 2$, Top-K	FID	-0.77924	$[-1.06333, -0.50527]$
$B = 2$, Top-K	sFID	-0.59955	$[-0.78191, -0.41850]$
$B = 2$, Top-K	Precision	-0.00166	$[-0.00474, 0.00142]$
$B = 2$, Top-K	Recall	$+0.01950$	$[0.01130, 0.02760]$
$B = 4$, Top-K	FID	-1.56024	$[-2.00764, -1.13764]$
$B = 4$, Top-K	sFID	-1.08012	$[-1.40476, -0.76193]$
$B = 4$, Top-K	Precision	$+0.00120$	$[-0.00262, 0.00520]$
$B = 4$, Top-K	Recall	$+0.02740$	$[0.02010, 0.03470]$
$B = 2$, Inception-DPP	FID	$+0.28836$	$[-0.00163, 0.57274]$
$B = 2$, Inception-DPP	sFID	-0.23635	$[-0.44258, -0.03172]$
$B = 2$, Inception-DPP	Precision	$+0.00530$	$[0.00192, 0.00876]$
$B = 2$, Inception-DPP	Recall	-0.01600	$[-0.02450, -0.00750]$
$B = 4$, Inception-DPP	FID	$+0.48984$	$[-0.01432, 0.97281]$
$B = 4$, Inception-DPP	sFID	-0.39770	$[-0.70978, -0.09281]$
$B = 4$, Inception-DPP	Precision	$+0.01844$	$[0.01388, 0.02312]$
$B = 4$, Inception-DPP	Recall	-0.01920	$[-0.02690, -0.01140]$

Evaluation and uncertainty. Both confirmations use the unmodified official ADM evaluator, reference statistics, and Inception graph. Retained membership indicators exactly replay the official precision and recall outputs. FID and sFID intervals use a first-order influence bootstrap over the same 1,000 generated classes in both policies, anchored at the official metric difference. Precision

resamples paired generated-class membership means with the real manifold fixed; recall resamples the same 10K real-query identities with both generated manifolds fixed. Each interval uses 20,000 paired draws and is unadjusted for multiple comparisons. The intervals describe composition sensitivity within each realized replication; regeneration, parameter selection, and manifold refitting are separate sources of uncertainty. IS is reported as a point estimate. Fresh generation supplies the independent replication evidence, and the intervals quantify its conditional metric uncertainty.

Artifacts. The `sit_retention_independent_20260922` and `dit_repeated_confirmation_20260922` bundles retain the frozen policies, independent seed definitions, class-balanced archives, and metric-specific interval records. The generated tables and anonymized source checksums are reproduced by `generate_final_confirmation_20260923.py -check`.

F Cross-Representation Audits

F.1 DINOv2 distribution metrics

Table 49: Post-hoc DINOv2 audit of SiT-original outputs, with DINO-DPP as the comparator. Lower DINO-FD and DINO-KID are better. Intervals are for each selector's KID difference from DINO-DPP.

B	Selector	DINO-FD	DINO-KID	ΔKID vs DPP	Paired 95% CI
2	Top-K	171.085	0.121742	-0.014438	$[-0.026295, -0.003827]$
2	DPP	181.241	0.136180	0	–
2	QUBO	177.164	0.131368	-0.004812	$[-0.016203, +0.008394]$
3	Top-K	161.270	0.100756	-0.011998	$[-0.020872, -0.004545]$
3	DPP	166.461	0.112754	0	–
3	QUBO	164.646	0.110950	-0.001804	$[-0.011727, +0.006513]$
4	Top-K	164.139	0.096118	-0.012412	$[-0.020433, -0.005366]$
4	DPP	165.811	0.108530	0	–
4	QUBO	163.344	0.103949	-0.004581	$[-0.012172, +0.002795]$

DINO-KID applies the KID estimator (Bińkowski et al., 2018) to DINO features, using the cubic kernel $((x^\top y/d) + 1)^3$. Paired intervals resample ImageNet-class blocks while using identical generated and reference indices. This representation audit reuses the frozen generated images.

F.2 CLIP-CMMD

CMMD (Jayasumana et al., 2024) uses L2-normalized OpenAI CLIP ViT-L/14@336px embeddings (Radford et al., 2021), Gaussian bandwidth 10, and scale 1000. The embedding implementation uses the official checkpoint and preprocessing but a PyTorch path rather than the original JAX/Scenic executable. All QUBO–DPP intervals in Table 50 cross zero; the later dynamic comparisons appear in Table 26.

G Qualitative Population Example

H Runtime and Memory

The nonmonotone QUBO timings reflect different runner implementations and shared-cluster load. We therefore ran the controlled stage-level profile in Table 52. In that implementation and shared-load profile, online features cost much more than swap search. At $B = 4$, adding the measured components gives approximately 91.63 s for Top-K and DINO-DPP, 149.09 s for matched Inception-DPP, and 149.17 s for Moment-QUBO. These are incremental one-round measurements. NFE matching controls denoiser calls; the profile separately measures feature-computation cost.

Table 50: CLIP-CMMD audit of SiT-original outputs, with DINO-DPP as the comparator. Lower is better. Intervals are for each selector's block-CMMD difference from DINO-DPP.

B	Selector	Full CMMD	Block Δ vs DPP	Paired 95% CI
2	Top-K	0.401674	-0.019124	$[-0.056334, +0.013970]$
2	DPP	0.420743	0	–
2	QUBO	0.424756	+0.004289	$[-0.033903, +0.036818]$
3	Top-K	0.342576	-0.024810	$[-0.054466, +0.001746]$
3	DPP	0.367918	0	–
3	QUBO	0.370395	+0.002301	$[-0.026355, +0.031285]$
4	Top-K	0.319269	-0.029607	$[-0.054073, -0.008395]$
4	DPP	0.349416	0	–
4	QUBO	0.347282	-0.002348	$[-0.033135, +0.022828]$

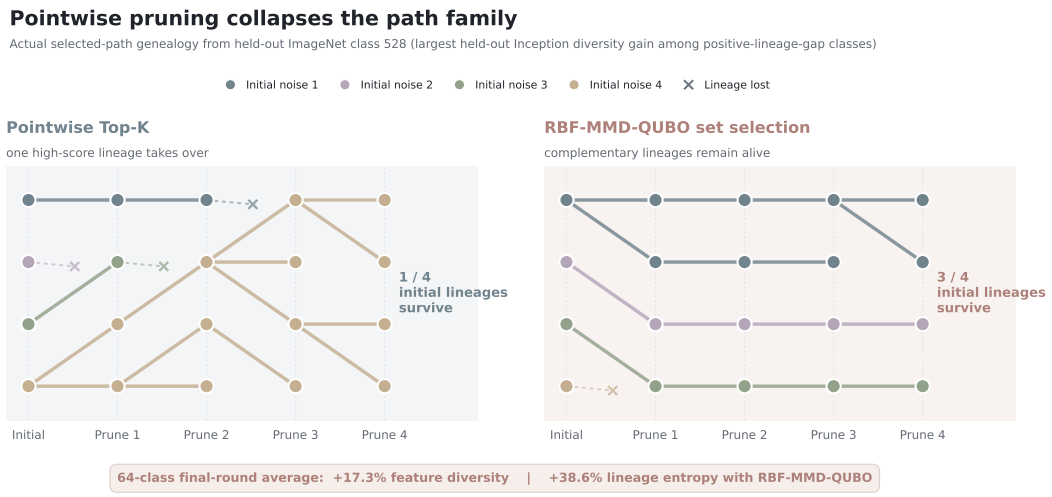

Figure 10: An actual Search-over-Paths genealogy for ImageNet class 528, $K = 4$, at equal compute. Nodes are candidate paths and thick colored edges indicate retained ancestry. Pointwise Top-K concentrates all final paths in one root family; RBF-MMD-QUBO preserves three of four. This is the RBF-kernel illustration, distinct from the main moment-kernel experiments.

H.1 Complete warm-model noise-to-image latency

We additionally profile the complete pipeline from fresh Gaussian noise to 50 final uint8 images, including the initial 222-NFE integration, all four branching/denoising rounds, clean predictions, VAE decoding/transfers, required features, and subset construction/optimization. Pure Heun-250 provides the no-search cost reference. Top-K computes DINO quality only; matched Inception-DPP and Moment-QUBO also compute Inception features. The existing SiT checkpoint, four-round schedule, $K = 50$, $\alpha = 0.56$, $\eta = 2$ and selectors are retained, using FP32 with TF32 enabled and eight CPU threads.

The three blocks use classes 7, 207 and 809 on separate NVIDIA H20-3e devices; methods run serially within each block in rotated order. One full maximum-shape QUBO warmup is discarded, with loading and warmup recorded separately. Every timed candidate is generated afresh. Model loading, artifact writes and ADM evaluation are outside its latency boundary. All networks remain resident even for the base sampler, so peak allocation describes the shared profiling process. Measurements include contemporaneous shared-device load. An initial warmup with an incomplete alpha schedule is archived separately from measured rows.

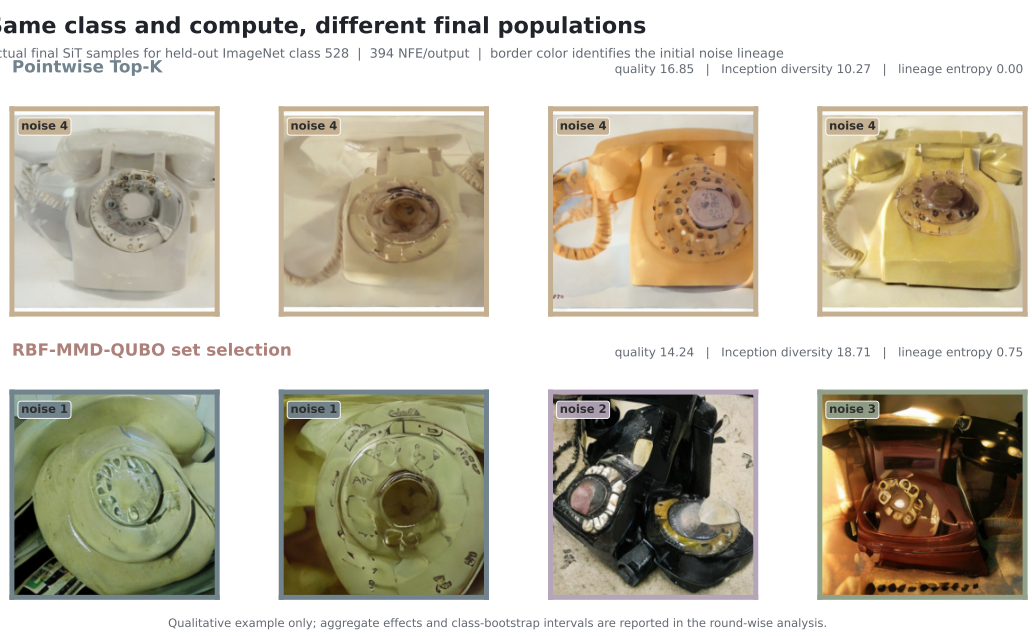

Figure 11: Final samples corresponding to the genealogy in Figure 10. Border colors denote initial noise lineages. Both methods use 394 NFE/output. This class was selected by the objective rule of maximum held-out Inception diversity gain among classes for which RBF-MMD-QUBO retains more final lineages.

Table 51: Median selector time per pruning round from the formal SiT traces.

B	M	Top-K (ms)	DPP (ms)	QUBO (ms)
2	100	0.100	1.071	466.561
3	150	0.124	1.277	416.505
4	200	0.233	12.952	41.837

Figure 5 in the main paper pairs these separately profiled costs with the existing formal 50K quality measurements.

This small profile updates the cost evidence without rerunning the formal quality study. At $B = 4$, within-block QUBO/Top-K latency ratios range from 1.001 to 1.018 (median 1.005). QUBO construction plus optimization takes a median 0.0448 s across all four rounds, or 0.0234% of complete latency; median initial integration and branch denoising alone take 93.84 and 67.17 s. The resident-process peak is 54.53 GiB for all three $B = 4$ selectors. This complete pipeline and the earlier one-round profile use different implementations and loads. Thread control and computing only required features accompany the complete-pipeline run; its measurements describe their combined execution on the GPU.

I Solver and CIM Experiments

Physical CIM platforms implement Ising-coupled optical dynamics (Inagaki et al., 2016; Wei et al., 2026); the experiments below distinguish software solvers, numerical simulation and supplied device returns.

Ten DiT class matrices were frozen before annealing evaluation: two development matrices, four held-out matrices with the largest greedy-to-one-swap gap, and four held-out matrices chosen by a fixed SHA-256 rank. Every instance has $M = 200$, $K = 50$ and is rescored using the original float64 QUBO.

Table 52: Controlled one-round SiT stage profile. Absolute values are hardware-specific and were collected under shared load.

B	M	Branch denoise (s)	Feature pipeline (s)	Swap (s)	Peak GiB
1	50	6.397	28.441	0.4677	16.64
2	100	12.236	61.554	0.3412	33.83
3	150	20.924	101.973	0.4218	45.93
4	200	19.645	129.412	0.1170	54.53

Table 53: Complete warm-model latency per 50 outputs. Three fixed class/device blocks; ranges span the three observed measurements. Relative cost is the median of the three within-block ratios to pure Heun-250.

B	Method	NFE/image	Median (s)	Block range (s)	Relative cost
0	Base Heun-250	250	107.70	66.80–109.05	1.000
1	Top-K	265	120.27	74.03–123.31	1.117
2	Top-K	308	139.82	88.57–146.39	1.326
2	Inception-DPP	308	144.30	89.58–146.44	1.341
2	Moment-QUBO	308	144.36	89.61–145.37	1.340
4	Top-K	394	192.06	118.73–194.66	1.783
4	Inception-DPP	394	194.39	120.83–195.39	1.809
4	Moment-QUBO	394	193.06	120.83–194.85	1.793

The fixed-K control, historically labeled FSA, uses uniform one-out/one-in proposals, Metropolis acceptance and inverse-linear cooling $T_t = T_0/(t+1)$. Its cooling schedule follows fast annealing (Szu & Hartley, 1987), with discrete swaps replacing continuous Cauchy proposals. It uses 128 reads, 5,000 proposed swaps per read, and temperature scale 1.0. It improves all four hard held-out matrices and one of four hash-random matrices, but every best state is exactly the one-swap state. A subsequent gate evaluates the complete two-swap neighborhood and FSA up to 512 reads and 20,000 sweeps; neither beats one-swap on any of the ten matrices.

The numerical SimCIM control simulates coherent-Ising-machine dynamics (Tiunov et al., 2019); its outputs in this test fail exact cardinality. The physical-device experiments below use separate instances.

For physical evaluation, each Ising instance has also been exported as a 201×201 Kaiwu complete matrix with one auxiliary spin and signed int14 couplings. A separate precision-split package expands each instance to 1,995–2,206 hardware variables. This expansion trades bits for coefficient precision while retaining 200 logical candidate variables. Before opening any new hardware return, we registered an encoding-fidelity screen on the union of the greedy and one-swap anchors and their complete exact-K one-out/one-in neighborhoods. The direct int14 encoding preserves 0.702 mean sampled pair ordering and reaches 3.56×10^{-3} worst float64 regret. Precision splitting preserves the greedy-to-one-swap direction on all 10 matrices, obtains 0.9981 mean pair-order agreement (minimum 0.9967), and limits worst float64 regret to 1.24×10^{-6}, passing all five preregistered encoding checks. Thus the precision-split encoding passes the software fidelity screen, while the compact direct encoding provides a diagnostic baseline.

The supplied reference states define the best-known classical energy targets. The frozen physical gate requires strict split-copy consistency, exact $K = 50$ without voting or cardinality repair, at least four of eight held-out matrices beating constructive greedy under the original float64 QUBO, and a negative mean held-out energy difference. These ten class-local instances complete the software encoding screen; the physical returns below concern the two corrected hardware-scale move-QUBOs and their historical predecessors.

Table 54: Solver results on the eight held-out DiT matrices. Four matrices have the largest held-out greedy-to-one-swap gaps; four use fixed hash ranks.

Solver	Exact-K	Beats greedy	Beats one-swap	Interpretation
Penalty GPU-SA	8/8	0/8	0/8	feasible states worse than greedy
Fixed-K FSA	8/8	5/8	0/8	recovers one-swap improvements
Numerical SimCIM	0/8	0/8	0/8	collapses to wrong cardinality

I.1 Physical execution with the corrected quality floor

Two new signed-int14 move-QUBOs enforce the original per-class quality condition on frozen 200K-candidate SiT pools:

$$|S_c| = 50, \qquad \sum_{i \in S_c} q_i \geq 0.995 \sum_{i \in \mathrm{TopK}_c} q_i. \tag{41}$$

Here q_i is the frozen raw DINO target-class score. The development and independent instances contain 5,912 and 5,770 exchange variables, or 5,913 and 5,771 spins including the gauge. They restrict the global real-reference moment objective in Equation 45 to disjoint within-class exchanges. These two corrected instances supply the quality-constrained physical execution results below; the following subsection retains the historical 6,858/6,615-spin instances and their weaker implemented threshold.

For each class, let Q_{0c} be its initial selected quality sum and Δq_p the quality change of exchange p. Disjoint removal and addition endpoints preserve cardinality for every bit vector. The constructor additionally requires

$$Q_{0c} + \sum_{p \in \mathcal{P}_c} \min(\Delta q_p, 0) \geq 0.995 \sum_{i \in \mathrm{TopK}_c} q_i. \tag{42}$$

The left side is the exact worst quality sum over the exchange hypercube, so the entire binary domain satisfies the floor. Independent checks against all 2,000 source feature files give worst-case class retention ratios 0.99500000187 and 0.99500000451. Coupling quantization changes the objective coefficients while preserving this feasible domain.

Table 55: Corrected signed-int14 physical instances. Each of the 100 supplied ranked states per instance attains the long-budget raw Kaiwu SA best-known energy and satisfies both exact class cardinality and Equation 41. Counts describe the supplied files. Minimum retention is over all returned states and classes.

Pool	Spins	Best-known H	CIM H	Valid returns	Unique selections	Min. retention
Development	5,913	$-9,898,870$	$-9,898,870$	100/100	23	0.995000619276
Independent	5,771	$-10,883,280$	$-10,883,280$	100/100	9	0.995000087694

The canonical objective is $H(s) = s^\top W^{(14)} s$, minimized on the full symmetric matrix. For the device convention $E_{\mathrm{device}} = -s^\top J s$, the corresponding device input is $J = -W^{(14)}$. Independent int64 sparse quadratic forms and twice the upper-triangle sum agree on all 200 returned states; global spin flips preserve both energy and decoded selection. The lowest integer energy, with ties resolved by earliest returned rank, fixes rank 1 as the primary state in each pool before downstream evaluation. Raw spins are decoded directly by $y_i = (s_i g + 1)/2$.

Raw solver quality at a 10 ms budget. We evaluate ordinary `kaiwu.classical.SimulatedAnnealingOptimizer` on these same corrected integer matrices. The SA clock runs from the native solve call to return of its complete native solution pool, including internal initialization. Matrix loading, initial-spin generation, optimizer construction, offline rescoring and artifact writes are outside this clock. Each call starts independently. All native pool states are retained unchanged; integer rescoring and choosing the lowest energy pool member do not modify any spin. The energy summary includes only nonempty native pools returned within 10 ms. Late returns and calls with no native output are counted separately. Both solver records use their original states without external refinement. The CIM duration of 10 ms is supplied by the

result submitter in the current update; the returned files contain ranked states. Table 56 reports energies, and Table 57 records the SA call accounting. Development uses initial temperature 12000,

Table 56: Raw solver quality at the 10 ms budget on the corrected 5,913/5,771-spin matrices. SA energy statistics summarize the minimum-energy member of each native pool returned on time. CIM statistics summarize the supplied ranked states, with time attributed to the submitter. H_{ref} is the long-budget best-known energy in Table 55. All energies minimize the same full-matrix Hamiltonian; spins are used as returned.

Pool	Solver	Recorded states/calls	Best H	Median H	Best $H - H_{\mathrm{ref}}$
Development	CIM	100 supplied	$-9{,}898{,}870$	$-9{,}898{,}870$	0
Development	Native Kaiwu SA	96/100 calls on time	$-440{,}654$	$-40{,}216$	9,458,216
Independent	CIM	100 supplied	$-10{,}883{,}280$	$-10{,}883{,}280$	0
Independent	Native Kaiwu SA	98/100 calls on time	$-632{,}732$	$-40{,}254$	10,250,548

Table 57: Native-SA call accounting for the corrected-matrix 10 ms comparison. On-time, late, and no-output categories partition all attempted calls. The API median uses all 100 calls; maximum on-time duration uses eligible nonempty calls. Target hits count eligible calls reaching H_{ref}.

Pool	On time	Late	No output	Median API (ms)	Max. on time (ms)	Target hits
Development	96	4	0	8.3456	9.8441	0/96
Independent	98	2	0	8.0813	9.9144	0/98

cutoff 0.5, cooling factor 0.729656404313, 1 single-spin proposal per temperature level, native pool capacity 100, and 1 SA process. Independent uses initial temperature 12000, cutoff 0.5, cooling factor 0.729656404313, 1 single-spin proposal per temperature level, native pool capacity 100, and 1 SA process. The parameters are selected on separate calibration seeds before the 100 formal calls per matrix; BLAS uses 16 threads. The native-SA parameters, source hashes, original solution pools and timing records are archived with the `cim_corrected_sa_10ms_20260926` study. This experiment measures raw integer-solution quality; the following ADM results evaluate the previously frozen long-budget reference selections.

CIM timing scope (per shot or batch, and inclusion of programming/readout) was not recorded in the supplied files.

Long-budget reference and decoded-image evaluation. The best-known reference is the best raw ordinary Kaiwu SA state on the same integer matrix. Four independent cold-start SDK calls use initial temperature 12,000, cutoff 0.5, cooling factor 0.9949697895411733, 2,001 temperature levels, and 100 single-spin proposals per level. The SDK retains up to 100 states per call; raw states pass the same decoding and quality checks. No external swap or polish is applied to either solver's states. All four calls per instance reach the listed reference, with median API durations 7.354/6.809 s. These calls define the long-budget reference for the following decoded-image evaluation. The 10 ms comparison in Table 56 uses separate SA calls.

Table 58: Corrected CIM versus the long-budget raw Kaiwu SA reference: four official ADM-50K evaluations assembled from cached candidate images. All selections satisfy Equation 41; additional diffusion NFE is zero. Entries are point estimates for the fixed primary states.

Pool	State	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
Development	CIM, raw rank 1	7.943551	8.662544	66.9301	0.86310	0.3752
Development	Kaiwu SA, best raw	7.943611	8.662290	66.9300	0.86310	0.3749
Independent	CIM, raw rank 1	8.080106	9.013176	66.7941	0.86100	0.3734
Independent	Kaiwu SA, best raw	8.080297	9.012702	66.7958	0.86092	0.3733

Each primary CIM selection differs from its long-budget raw SA reference in three exchange bits, replacing three of 50,000 selected images. Their float QUBO differences are

Table 59: Historical physical execution under the weaker implemented quality threshold: two signed-int14 population-selection instances. Spins include the gauge variable. All 100 supplied ranked states attain the listed CIM energy and retain exactly 50 images per class. Feasibility here refers to class cardinality. The frozen classical energies are best-known references. $H(v) = v^\top W^{(14)} v$ is minimized.

Pool	Spins	Classical reference H	CIM H	Feasible returns
Development	6,858	$-14{,}787{,}206$	$-14{,}787{,}206$	100/100
Independent	6,615	$-13{,}356{,}620$	$-13{,}356{,}620$	100/100

-6.13×10^{-9} and $+2.41 \times 10^{-9}$, despite equal integer energies. CIM-minus-reference FID is $-0.0000594/-0.0001905$, with exactly replayed ADM precision/recall memberships. These results complete the execution chain from quality-constrained population selection to physical spins and decoded image populations. Protocols, matrix checksums, raw states, and downstream records are retained in the `cim_quality_corrected_return_20260922` and `cim_quality_corrected_downstream_20260922` bundles.

I.2 Historical hardware-scale coordinated-move QUBOs

The physical execution interface restricts a frozen population problem to class-preserving moves. Ordinary simulated annealing (Kirkpatrick et al., 1983) provides the classical comparison on the same integer instances:

frozen population $\longrightarrow$ feasible joint moves $\longrightarrow$ int14 Ising $\longrightarrow$ decoded selection.

The $M = 200$ class-local matrices above come from the single-pruning DINO RBF-MMD control. We also construct two much larger QUBOs for global refinement of a frozen 200K-candidate SiT pool. Starting from a feasible 50K selection, each binary variable activates a concrete one-out/one-in move. Within each class, disjoint move endpoints preserve exactly 50 images. The archived constructor applies a weaker quality threshold than intended, as audited below. Expanding the original normalized first- and second-moment objective over these moves gives its exact quadratic restriction, without a cardinality or quality penalty.

I.2.1 Exact restriction of the global moment objective

Let $N = 50{,}000$ be the total retained population, let $u_i = f_i - c$ be an Inception feature centered by the real-reference mean, and define

$$\mu(S) = \frac{1}{N} \sum_{i \in S} u_i, \tag{43}$$

$$D(S) = \frac{1}{N} \sum_{i \in S} u_i u_i^\top - C_{\mathrm{ref}}, \tag{44}$$

$$\mathcal{H}(S) = \frac{\|\mu(S)\|_2^2}{\tau_1} + \frac{\|D(S)\|_F^2}{\tau_2}. \tag{45}$$

Here C_{ref} is the real-reference covariance, $\tau_1 = \max\{\mathrm{tr}\, C_{\mathrm{ref}}, 10^{-12}\}$, and $\tau_2 = \max\{\tau_1^2 + 2\|C_{\mathrm{ref}}\|_F^2, 10^{-12}\}$. This global hardware experiment accesses real-reference moments; it is distinct from the candidate-self-reference dynamic search in the main curve. For a proposed move $p = (r_p \to a_p)$, define its normalized first- and second-moment changes

$$d_p = \frac{u_{a_p} - u_{r_p}}{N}, \qquad T_p = \frac{u_{a_p} u_{a_p}^\top - u_{r_p} u_{r_p}^\top}{N}. \tag{46}$$

Writing $\mu_0 = \mu(S_0)$ and $D_0 = D(S_0)$, a move vector $y \in \{0,1\}^P$ gives $\mu(y) = \mu_0 + \sum_p d_p y_p$ and $D(y) = D_0 + \sum_p T_p y_p$. Expanding the two squared norms and using $y_p^2 = y_p$ yields

$$\mathcal{H}(S(y)) - \mathcal{H}(S_0) = \sum_p \ell_p y_p + \sum_{p<q} Q_{pq} y_p y_q, \tag{47}$$

where

$$\ell_p = \frac{2\mu_0^\top d_p + \|d_p\|_2^2}{\tau_1} + \frac{2\langle D_0, T_p\rangle_F + \|T_p\|_F^2}{\tau_2}, \tag{48}$$

$$Q_{pq} = \frac{2d_p^\top d_q}{\tau_1} + \frac{2\langle T_p, T_q\rangle_F}{\tau_2}. \tag{49}$$

Equation 47 is an exact algebraic restriction of the same frozen moment objective. The initial selection reproduces its registered objective with zero measured error. For the development float reference, the stored move-QUBO predicts a change of -0.002847118639, and direct feature-based recomputation gives -0.002847117838, within the 2×10^{-8} audit tolerance.

I.2.2 Cardinality and the historical implemented quality threshold

For class c, let the initial feasible set S_{0c} contain 50 candidates, and let $\mathcal{P}_c$ index its moves. All move-index sets and sums in the following feasibility argument are restricted to $\mathcal{P}_c$. Each move removes $r_p \in S_{0c}$ and inserts $a_p \notin S_{0c}$. We require distinct remove endpoints and distinct add endpoints within the class. Therefore, for every bit vector y,

$$S_c(y) = (S_{0c} \setminus \{r_p : y_p = 1\}) \cup \{a_p : y_p = 1\} \quad \Longrightarrow \quad |S_c(y)| = 50. \tag{50}$$

The move construction also preserves its implemented threshold. Write $Q_{0c} = \sum_{i \in S_{0c}} q_i$ and $\Delta q_p = q_{a_p} - q_{r_p}$. Let L_c denote the threshold applied to the class-quality sum. One move per class is an anchor that may have negative Δq_p; it is admitted only when $Q_{0c} + \Delta q_p \geq L_c$. Every other move has $\Delta q_p \geq 0$. Hence

$$Q_c(y) = Q_{0c} + \sum_p y_p \Delta q_p \geq \min\{Q_{0c}, Q_{0c} + \Delta q_{\text{anchor}}\} \geq L_c. \tag{51}$$

Thus cardinality and the implemented threshold hold for every state in the binary hypercube through the move construction. We choose the anchor as the best individually floor-feasible move. The remaining moves come from a maximum-cardinality, minimum-linear-cost bipartite matching over quality-nondecreasing remove/add pairs. This gives 2–15 variables per class in the development pool and 2–9 in the independent pool.

Quality-threshold audit. The intended floor is $0.995 \sum_{i \in \mathrm{TopK}_c} q_i$, where TopK_c contains the 50 highest-scoring candidates under raw verifier quality. The archived constructor instead loads $L_c = 0.995\,\mathrm{mean}_{i \in \mathrm{TopK}_c}\, q_i$ and compares this mean-scale threshold with a class sum, weakening it by a factor of 50. The preceding guarantee therefore applies to the stored threshold. Recomputing quality from the frozen candidate archives confirms that the initial selections satisfy the intended floor, but each of these two archived move sets admits a violating state in each of its 1,000 classes. All 100 supplied CIM states in each pool preserve exact class cardinality; each violates the intended quality floor in 999/997 classes for development/independent. Their minimum class-quality retention ratios are 0.971766/0.971677. These instances test physical execution and solution quality on the same frozen integer Hamiltonian. Their downstream metrics describe the decoded populations under the weaker implemented constraint.

Table 60: Classical audit of the two historical hardware-scale signed-int14 move-QUBOs. $H(v) = v^\top W^{(14)} v$ is minimized on the full symmetric Kaiwu matrix; H_{SA} is the historical 384-read warm-SA+polish reference. Pair counts refer to nonzero upper-triangular logical couplings and exclude the auxiliary gauge fields.

Pool	Logical vars.	Matrix order	Pair couplings	H_{warm}	H_{SA}	Active moves
Development	6,857	6,858	416,058	$-14{,}787{,}206$	$-14{,}787{,}206$	2,880
Independent	6,614	6,615	360,059	$-13{,}356{,}620$	$-13{,}356{,}620$	2,799

I.2.3 Frozen-pool downstream scaling

The development move-QUBO reselects images from the same frozen 200K pool as the prospective full/linear/shuffled ablation, so it adds zero diffusion NFE. Table 61 reports the resulting ADM

metrics. The move construction and the 5,000-swap continuation were designed after the original 200-swap result was observed. Both also use the mean/sum-mismatched threshold identified above, whereas the original 200-swap selection satisfies the intended floor. These continuations change the constraint as well as the search depth or neighborhood. Direct rescoring of the 5,000-swap selection finds 890 classes below the intended floor, with minimum class-quality retention 0.593399 and mean class retention 0.949769.

Table 61: ImageNet-256 ADM evaluation on one frozen 200K-candidate SiT pool. All rows contain 50K class-balanced images. "Depth" uses algorithm-specific units: for move-QUBO it is the number of active coordinated moves, while for one-swap it is the accepted-swap cap. The move-QUBO and deep one-swap rows use the weaker implemented quality threshold described in Appendix I.2.2.

Selector	Depth	Time (s)	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
Pointwise Top-K	–	–	9.43659	8.82624	**67.47028**	**0.88576**	0.32990
Inception-DPP	–	–	7.95620	8.66995	66.82327	0.86038	0.38250
Global Moment-QUBO	200	153.14	7.91294	8.65794	66.81146	0.86010	0.38220
Coordinated move-QUBO	2,880	71.38	6.88247	8.50906	66.49413	0.85016	0.40230
Deep one-swap	5,000	470.32	**4.83768**	**8.03883**	64.68868	0.81244	**0.46860**

Relative to the 200-swap global QUBO, coordinated moves lower FID by 1.03047 and raise recall by 0.02010, while reducing precision by 0.00994. Deep one-swap lowers FID further but reaches its 5,000-swap cap before convergence. These shifts combine additional optimization with the relaxed quality constraint; they do not isolate the effect of solver depth.

The move-QUBO itself provides a more surgical structure check. Selecting all negative linear moves nearly matches the best-known full float-QUBO reference, as shown in Table 62. The full pairwise objective changes 186 move decisions: it deactivates 110 and activates 76, reducing the active count by 34. Its float-objective improvement over linear-negative selection is 3.89×10^{-6}. This objective comparison complements the downstream comparison across search neighborhoods and depths in Table 61.

Table 62: Objective decomposition on the development move-QUBO. All energies are rescored using the original float coefficients.

State	Active	Linear	Pairwise	Total $\Delta\mathcal{H}$
Negative-linear	2,914	-0.00336519	$+0.00052196$	-0.00284323
Float-QUBO reference	2,880	-0.00335958	$+0.00051246$	-0.00284712
Int14 reference	2,880	-0.00336441	$+0.00051906$	-0.00284535

Both matrices use one auxiliary gauge spin, are symmetric with zero diagonal, contain no saturated coefficients in the signed-int14 range, and fit below the 10K-variable hardware limit. Random-scan Metropolis SA uses 384 reads, 2,000 sweeps per read, a temperature schedule from 12,000 to 0.5, and 64 CPU threads. Its wrapper takes 9.24 s and 6.33 s on the development and independent instances, including problem preparation, compilation, SA, polish and audits, but excluding construction of the pre-existing warm reference. In both cases SA matches the warm-reference Ising energy; deterministic polishing accepts zero further flips. The decoded float-QUBO changes are -2.84535×10^{-3} and -2.81812×10^{-3}. These audits establish exact class cardinality, hardware-scale encoding, and reproducible classical energy thresholds.

I.2.4 Annealing transition and audit protocol

We fix the gauge to $g = +1$ during classical annealing; global-flip symmetry makes this equivalent to searching both gauge sectors. For the symmetric, zero-diagonal integer matrix $W^{(14)}$ and energy $H(s) = s^\top W^{(14)} s$, flipping logical spin s_i changes the energy by

$$\Delta_i H = -4 s_i \big(W^{(14)} s\big)_i. \tag{52}$$

Maintaining the local field $W^{(14)} s$ permits an incremental update over the nonzero neighbors of the flipped spin. At sweep t, we make as many proposals as logical spins, sampling each index

independently with replacement. An uphill move is accepted with probability

$$p_{\text{accept}} = \min\{1, \exp(-\Delta_i H/T_t)\}, \qquad T_t = 12000 \left(\frac{0.5}{12000}\right)^{t/1999}. \tag{53}$$

Each instance uses 384 reads, 2,000 sweeps, 64 CPU threads, and a frozen seed. The read allocation covers the warm reference, perturbed warm references, Bernoulli(0.42) states, and the all-zero logical state. The best state seen at any temperature is retained and then passed to deterministic steepest single-spin polish.

Every reported state passes five checks: incremental energy equals the integer bit-form energy, including its constant term; this energy equals direct $s^\top W^{(14)} s$; the matrix is symmetric; its diagonal is zero; and decoding followed by original float64 rescoring reproduces the recorded selection. The best read on both instances is the unmodified warm reference and polishing accepts zero flips. All 128 Bernoulli(0.42) cold reads in each archived log reach the target during annealing; their terminal target counts are 127/128 and 128/128. These subgroup energies precede polish. Only the overall winner's bits were archived, motivating the explicit raw-state control below. A physical solver is compared against the same integer Hamiltonian.

I.2.5 Custom C++ cold-start SA without postprocessing

We separate the historical best-known reference from a custom C++ SA control. This implementation is distinct from both ordinary Kaiwu SA below and FSA. Each matrix receives 100 independent Bernoulli(0.5) initial states, fixed gauge $g = +1$, 2,000 sweeps and the same 12,000-to-0.5 temperature schedule. Seeds are 20260914001/20260914002 and each pool uses 16 CPU threads. Each random initialization is updated solely by the prescribed Metropolis transitions. All initial, terminal and best-visited states are saved. Integer sparse-matrix rescoring and independent upper-triangle summation audit every saved state. Best-state tracking records visited states, and terminal and best-visited results are reported separately.

Table 63: Custom C++ SA and supplied physical returns on identical matrices. SA counts cover every attempted run in this control. CIM counts cover the supplied ranked top-100.

Pool	Target H	Raw SA terminal	SA best-visited	Supplied CIM
Development	$-14{,}787{,}206$	100/100	100/100	100/100
Independent	$-13{,}356{,}620$	100/100	100/100	100/100

Every custom C++ SA terminal state attains its target energy after cold-start annealing at this budget. The 100-read solver stages take 2.809/2.679 s; binary invocation including problem loading and state/table output takes 2.824/2.693 s. Python preparation and complete replay audits take 9.506/9.770 s in total, excluding the one-time binary compilation. These are shared-CPU batch timings for 100 reads.

For ADM evaluation, we select the custom C++ SA terminal state with the lowest integer energy, breaking ties by earliest read index (read 0 in both pools). This rule is fixed before evaluation. Compared with the historical reference, the selected states replace one image in the development pool and three in the independent pool; compared with CIM rank 1, each replaces 2 images. The move variables specify fixed legal exchanges that decoding applies directly. The common frozen 200K population and move subspace remain upstream of both solvers. This control supplies independently initialized raw classical states for the same integer problem.

I.2.6 Ordinary Kaiwu simulated annealing

We additionally run Kaiwu SDK 1.0.33a ordinary simulated annealing (QBOSON, 2026) (`SimulatedAnnealingOptimizer`). Each call starts from an independent unbiased random full spin vector, including the gauge spin. The unmodified SDK retains a pool of low-energy states and resets its working state to the best state after a temperature with accepted moves. We evaluate these native SDK outputs directly. The long-budget schedule uses $T_0 = 12000$, $T_{\min} = 0.5$, cooling multiplier 0.9949697895411733, and 2,001 temperature levels. Each temperature proposes one single-spin move per full matrix dimension (6,858/6,615 proposals). The native pool limit is 100;

Table 64: Ordinary Kaiwu SA on the two frozen signed-int14 matrices. Target counts refer to independent calls. Time is per-call API wall time under the stated parallel CPU load.

Pool	Best H	Target calls	Median (s)	Range (s)
Development	$-14{,}787{,}206$	100/100	334.122	330.530–497.711
Independent	$-13{,}356{,}620$	100/100	320.102	314.559–480.215

each matrix receives 100 independent calls. We use 32 external CPU workers, with SDK process count 1 and 1 BLAS thread per worker. Each call is an independent trial; its returned pool contains retained low-energy states from that trajectory. Table 64 reports call-level target counts and observed native API wall time; the timing includes the SDK initialization. The complete parallel batch takes 2282.930 s.

I.2.7 Ordinary SA at a short time budget

We also record 100 independent cold-start calls per matrix and clock with a 10 ms budget. The annealing clock starts after the initial dense-energy computation. The API clock starts immediately before the SDK solve call and includes that computation; the matrix and random initial spins are already resident. An observation wrapper saves the lowest-energy visited state whose copy completed before the deadline, or the preavailable random initial state if none did. Observation and copying count against the selected clock. The wrapper preserves the SDK random proposals, Metropolis acceptance, and best-state reset. We report the saved deadline checkpoints and unmodified native API returns separately. A separate unmodified-SDK control keeps the minimum-energy returned pool state only when the native API returns within the budget, otherwise retaining the preavailable initial state. Table 65 distinguishes these outputs. The hardware-return submitter reports a CIM duration of 10 ms; its per-shot versus batch scope and initialization, programming, and readout breakdown remain unspecified. The CIM rows summarize the supplied ranked returns, whereas the SA rows include all registered independent attempts.

Short-budget parameter selection. The temperature endpoints and native pool size remain as above. Short-budget calls execute serially on an Intel Xeon Platinum 8558 CPU, using 16 BLAS threads per call for both the checkpoint and native API controls. For nominal proposal count L and u proposals per temperature, the geometric cooling factor is $\rho = (0.5/12000)^{1/\max(1,\lfloor L/u \rfloor)}$. The checkpoint calibration uses 8 paired initial seeds, nominal proposal counts $32, 64, 128, 256, 512, 1024$ and proposals per temperature $1, 4, 16$. Each matrix/clock chooses the lowest mean integer energy, breaking ties by fewer nominal proposals and then fewer proposals per temperature. For the annealing checkpoint, development/independent use nominal proposals 128/128 and 1/4 proposals per temperature. For the API checkpoint, development/independent use nominal proposals 64/128 and 1/4 proposals per temperature. The native API calibration uses 8 paired seeds, nominal proposal counts $16, 32, 64, 128, 256$ and proposals per temperature $1, 4, 16$; late returns are replaced by the initial state when computing the calibration mean. Both matrices select 16 nominal proposals and 4 proposals per temperature. All 100/100 native formal calls return within the budget; observed median API times are 7.507/7.914 ms. For the annealing-checkpoint runs, the observed median complete API times are 14.794/13.685 ms; the retained states finish copying before the annealing deadline.

I.2.8 Supplied CIM and short-budget SA: downstream evaluation

Before new ADM evaluation, we fix the SA primary to independent call index 0 in the annealing-checkpoint experiment for each pool. This rule selects one predetermined attempt per pool. The CIM primary remains the previously frozen rank 1. The SA primary states finished copying at 9.999916/9.929113 ms. Every decoded archive contains 50,000 cached images, exactly 50 per class. Appendix I.2.2 audits the quality-threshold units. We decode the retained states directly and use the same official ADM evaluator, graph and reference; previously evaluated CIM metrics are reused only after verifying selection and source identities. Additional diffusion NFE is 0.

Table 65: Short-budget integer energies on identical frozen matrices, with the output and clock definitions in Section I.2.7. SA budgets are 10 ms under the indicated clock. CIM 10 ms is submitter-reported with timing scope unspecified. Counts are at the frozen target energy. Denominators are supplied ranked returns for CIM and all registered independent attempts for SA.

Pool	Output	Best H	Median H	At target
Development	Supplied CIM	$-14{,}787{,}206$	$-14{,}787{,}206$	100/100
Development	SA anneal checkpoint	$-1{,}046{,}486$	$-250{,}544$	0/100
Development	SA API checkpoint	$-971{,}038$	$-131{,}342$	0/100
Development	SA native API return	$-829{,}234$	$-162{,}874$	0/100
Independent	Supplied CIM	$-13{,}356{,}620$	$-13{,}356{,}620$	100/100
Independent	SA anneal checkpoint	$-942{,}984$	$-157{,}846$	0/100
Independent	SA API checkpoint	$-983{,}068$	$-77{,}596$	0/100
Independent	SA native API return	$-753{,}888$	$-89{,}502$	0/100

Table 66: Official ADM-50K metrics for the directly decoded frozen selections: rank 1 of each supplied CIM file and call 0 of each SA annealing-checkpoint experiment. The SA state comes from the predetermined first 10 ms annealing attempt in each pool.

Pool	Selection	H	FID	sFID	IS	Precision	Recall
Development	CIM rank 1	$-14{,}787{,}206$	6.8851	8.5260	66.4992	0.85070	0.40390
Development	SA 10 ms call 0	$-68{,}078$	7.5997	8.6060	66.7667	0.86134	0.38480
Independent	CIM rank 1	$-13{,}356{,}620$	7.0131	8.8775	66.3816	0.84756	0.40200
Independent	SA 10 ms call 0	$252{,}304$	7.7255	8.9846	66.6122	0.85776	0.38360

I.2.9 Physical return audit and timing protocol

The supplied hardware files contain 100 ranked states for each matrix. Direct int64 evaluation of $s^\top W^{(14)} s$, independently checked by upper-triangle summation, equals the reported energy in all 200 states: $-14{,}787{,}206$ for development and $-13{,}356{,}620$ for the independent pool. The earlier return optimized the opposite sign; the adapter $J = -W^{(14)}$ and new returned energies are consistent with the device convention $E_{\text{device}} = -s^\top J s$.

The development pool has 8 unique gauge-decoded selections, differing from the frozen SA bits in 0–3 moves. The independent pool has 38 unique selections, differing in 1–6 moves. All preserve 50 images in each of 1,000 classes. Appendix I.2.2 reports the mean/sum threshold audit for these historical instances. The corrected 5,913/5,771-spin results appear in Appendix I.1. Our audit directly decodes the returned states and rescores them. Rescoring the stored float QUBO gives CIM-minus-SA differences in $[-3.77 \times 10^{-8}, 3.74 \times 10^{-9}]$ and $[-1.36 \times 10^{-8}, 2.45 \times 10^{-8}]$; equal quantized energies coexist with distinct image selections and FID values. For ADM evaluation, we select the returned state with the lowest integer energy, breaking ties by earliest returned rank (rank 1 in both pools). This rule was fixed before assembling any downstream archive. We evaluate each primary selection and its historical same-int14 warm-SA-plus-polish reference with the frozen official ADM evaluator (Table 67). Each archive contains 50,000 cached images, exactly 50 per class; additional diffusion NFE is zero. The comparator uses the same int14 Hamiltonian. Source-image and selection hashes are checked during assembly.

CIM-minus-historical-reference FID is $-0.000210/-0.000158$. These descriptive differences show closely matched downstream quality on the two instances. The four evaluations use the same frozen evaluator, reference and graph on separate GPUs of the same model.

The original independent protocol requires strict $H_{\text{CIM}} < H_{\text{classical}}$; the observed equal energies miss this gate. The supplied historical record contains ranked states, energies, and a reported hardware origin; its submitter subsequently reported a 10 ms duration. Per-attempt initialization, upstream classical processing, full attempt counts, top-100 filtering, upload logs and timing components

Table 67: Historical weaker-floor physical returns versus the same-int14 warm-SA-plus-polish reference, distinct from the raw corrected-instance comparison in Table 58. All rows use official ImageNet-256 ADM-50K evaluation.

Pool	Selection	FID	sFID	IS	Precision	Recall
Development	CIM rank 1	6.885132	8.526043	66.499184	0.85070	0.40390
Development	Warm SA+polish	6.885343	8.525905	66.498550	0.85072	0.40390
Independent	CIM rank 1	7.013149	8.877548	66.381622	0.84756	0.40200
Independent	Warm SA+polish	7.013306	8.878097	66.381897	0.84762	0.40130

are outside the supplied record. The 100/100 counts use the ranked files as their denominators; Appendix I.2.6 specifies the clocks used for the 10 ms comparisons.

Future timed runs must retain all attempts, initialization and device/upload records under the frozen checksum and gauge convention. Only then can we estimate an independent-run target probability p_{hit} and

$$\mathrm{TTS}_{99} = t_{\mathrm{run}} \left\lceil \frac{\log(1-0.99)}{\log(1-p_{\mathrm{hit}})} \right\rceil, \qquad 0 < p_{\mathrm{hit}} < 1. \tag{54}$$

If $p_{\mathrm{hit}} = 1$, $\mathrm{TTS}_{99} = t_{\mathrm{run}}$; if $p_{\mathrm{hit}} = 0$, it is reported as unresolved at the observed read budget rather than extrapolated.

For future timed runs, the protocol separates programming/upload, optical evolution or annealing, readout, decoding, and total host-to-host latency. It specifies solver-only and I/O-inclusive TTS, the number of reads, and device batching. The completed downstream check compares decoded CIM states with a classical solution of equal integer energy and reports the resulting ADM metrics. The timing protocol aligns target energies, initialization, attempt accounting and latency components across devices.

J Precision Calibration and Complete-Policy Cost

This study separates two practical questions: how population objectives compare at similar ADM precision, and how their operating points compare with ordinary sampling when guidance and inference cost are included. The calibrated DiT comparison extends the fixed-policy transfer in Appendix E.2. Its DPP weight is selected using precision on development classes, and its final metrics use newly generated 50K populations. A subsequent ordinary-sampling evaluation supplies an additional cost–quality operating point.

J.1 Calibration and evaluation sequence

The calibration protocol was frozen on September 24 before generating its development populations. Throughout, Moment-QUBO keeps $\alpha = 0.56$ and $w_1 = w_2 = 1$, with four retention rounds and $K = 50$. The 32-class screen evaluates 24 configurations, producing 38,400 images. For SiT, the nine settings are ordinary sampling at CFG $\{1, 1.5, 2, 3, 5\}$, Moment-QUBO at $B \in \{2, 4\}$, and Top-K and Inception-DPP with $\eta = 4$ at $B = 4$. For DiT, the 15 settings are DDIM25 at CFG $\{1.5, 2, 3, 5, 8\}$ and, at each $B \in \{2, 4\}$, Moment-QUBO, Top-K, and Inception-DPP with $\eta \in \{4, 8, 16\}$. Only ADM precision enters calibration; FID and recall are evaluated on the complete 50K populations.

For each Moment-QUBO anchor, the screen selects the ordinary-sampling CFG and available same-budget DPP setting with the closest precision. Ties use the smaller numeric parameter, then the method name. The seven resulting pairs contain ten distinct configurations, which are evaluated without retuning on 64 disjoint classes and fresh noise (32,000 images). The prespecified advancement rule is an absolute precision point gap of at most 0.005. Table 68 reports every selected pair: only DiT $B = 2$ with DPP $\eta = 16$ advances. Its held-out precision interval is wider than the matching margin; advancement uses the registered point criterion. The other six pairs retain their observed precision gaps.

Table 68: All seven held-out calibration pairs, with 3,200 images per configuration. ΔP is Moment-QUBO minus its comparator. The gate uses $|\Delta P| \leq 0.005$; only the passing pair advances to formal 50K evaluation in the original protocol.

Anchor	Comparator	P_{Q}	P_{comp}	ΔP	Gate
SiT, $B = 2$	Base CFG $= 1.5$	0.81594	0.80156	$+0.014375$	Fail
SiT, $B = 4$	Base CFG $= 2$	0.84219	0.86031	-0.018125	Fail
SiT, $B = 4$	DPP $\eta = 4$	0.84219	0.85062	-0.008437	Fail
DiT, $B = 2$	Base CFG $= 3$	0.94156	0.92125	$+0.020312$	Fail
DiT, $B = 2$	DPP $\eta = 16$	0.94156	0.94094	$+0.000625$	Pass
DiT, $B = 4$	Base CFG $= 3$	0.93469	0.92125	$+0.013437$	Fail
DiT, $B = 4$	DPP $\eta = 8$	0.93469	0.92844	$+0.006250$	Fail

The formal comparison generates 50 images in each of all 1,000 ImageNet classes per endpoint, totaling 100K new images. Its initial and branch noise namespaces are separate from the screen, held-out stage, and earlier DiT experiments. The two policies share initial states and branch random numbers within this formal run, while their retained parents and subsequent candidates may diverge. Both use DiT-XL/2, CFG 1.5, $B = 2$, and the same Inception representation and DINO quality verifier.

On September 25, after these ADM results were known, we registered a follow-up consisting of conditional uncertainty, a CLIP evaluation of the two frozen populations, and 50K ordinary DDIM25 outputs at CFG 3. CFG 3 had already been selected by the screen; its original held-out precision gap remains outside the matching margin. The follow-up uses the same formal initial-noise class/slot indices. Retention changes ancestry, so a final output rank does not define an image-level pair. The combined studies generate 220,400 images, excluding implementation smoke checks; only the CFG 3 population is newly generated in the follow-up.

J.2 Formal metrics and conditional uncertainty

Table 69 reports all five official ADM metrics. Moment-QUBO has 0.77821 lower FID than calibrated DPP (6.0%), with precision difference 0.00004 and recall gain 0.01460. The conditional precision interval, $[-0.00340, 0.00352]$, lies inside the prespecified ± 0.005 margin. FID, sFID, and recall intervals favor Moment-QUBO at this matched operating point, whereas the IS point estimate is slightly higher for DPP.

Table 69: Complete DiT ImageNet-50K evaluations. Each row contains 1,000 classes with 50 outputs each. Seconds/50 is the median observed warm inference time for a complete class population. The two search policies use $B = 2$ and CFG 1.5; ordinary DDIM25 uses CFG 3.

Method	FID↓	sFID↓	IS↑	Prec.	Rec.	Seconds/50
Moment-QUBO ($B = 2$)	12.1624	9.4953	71.36	0.93484	0.2380	64.78
Inception-DPP ($\eta = 16$)	12.9406	9.8357	71.55	0.93480	0.2234	64.80
DDIM25 (CFG $= 3$)	9.5438	8.9941	72.13	0.91232	0.3685	23.96

All intervals in Table 70 use 10,000 paired resamples and are conditional on the realized populations and fixed reference. FID and sFID use a first-order class-cluster influence approximation, anchored at the official metric differences. The plug-in values agree with the official values within the registered 0.001 tolerance. Precision resamples 1,000 paired generated-class membership means with the real manifold fixed. Recall instead resamples the same 10,000 real-query identities, keeping both generated manifolds fixed. The saved memberships exactly replay the official precision and recall. These marginal 95% intervals measure class/query composition sensitivity; independent generation, manifold refitting, and parameter selection remain separate uncertainty sources.

The CLIP evaluation uses the frozen OpenAI ViT-L/14@336px encoder, L2-normalized 768-dimensional image features, and the same 10K real reference as the earlier representation audit. We compute biased Gaussian-kernel MMD with bandwidth $\sigma = 10$ and scale 1,000, using all 50K outputs per method. Exact class-pair kernel means support paired class-composition resampling

Table 70: Moment-QUBO minus calibrated Inception-DPP on their frozen DiT $B = 2$ 50K populations. FID/sFID, precision, recall, and CLIP-CMMD use the metric-specific conditional resampling procedures described in the text. Intervals are marginal 95% intervals.

Metric	Difference	Conditional 95% interval
FID	-0.7782	$[-1.1604, -0.4198]$
sFID	-0.3404	$[-0.5460, -0.1362]$
Precision	$+0.00004$	$[-0.00340, 0.00352]$
Recall	$+0.01460$	$[0.00600, 0.02320]$
CLIP-CMMD	$+0.001401$	$[-0.000542, 0.003275]$

with the real reference fixed. CLIP-CMMD is 0.194500 for Moment-QUBO and 0.193099 for DPP. The difference 0.001401 has interval $[-0.000542, 0.003275]$: DPP has the lower point estimate, while this interval leaves their CLIP ordering unresolved. The resolved improvements here concern Inception-space FID, sFID, and ADM recall.

J.3 Ordinary sampling and measured inference cost

Ordinary DDIM25 at CFG 3 achieves FID 9.54380, recall 0.36850, and precision 0.91232. Its precision is 2.252 percentage points below Moment-QUBO, while its FID, recall, and observed inference time are better. Thus it supplies a lower-precision operating point with a different cost–quality balance. Together with the matched DPP pair, these results locate the benefit of moment retention within high-precision population search and expose the trade-off with ordinary sampling. The measured points are reported individually, without interpolation to an unmeasured precision or time budget.

Inference timing starts from fresh Gaussian noise in an already loaded and warmed process and ends at the final CPU uint8 images. It includes denoising, required verifier and population-feature extraction, selection, and decoding. Model loading, warmup, archive writing, and offline metric evaluation are recorded separately. Each policy executes independently, with no shared-prefix or candidate reuse for timing. Counting both conditional and unconditional network evaluations, the two search policies require 112 image-level forwards per final output, compared with 50 for ordinary DDIM25.

The observed medians are 64.78, 64.80, and 23.96 seconds per 50 outputs for Moment-QUBO, DPP, and CFG 3, respectively. Devices share other workloads, and per-device medians vary from approximately 39–92 seconds for search and 23–34 seconds for CFG 3. The figures therefore describe complete-policy latency under the recorded resource conditions, rather than an isolated accelerator speed ratio. Device-stratified counts and timing distributions accompany the source records.

The accompanying artifact retains the registered protocols, class lists, noise definitions, population checksums, metric records and timing distributions. Its table generator binds the reported values to these records.

K Reproducibility and Audit Checklist

The code artifact includes:

- QUBO construction for quality–diversity, RBF-MMD, and moment matching;
- fixed-cardinality greedy, one-swap, two-swap, historical FSA controls, and ordinary Kaiwu SA with raw and time-budgeted returns;
- SiT and DiT dynamic path-search runners with deterministic seed namespaces and resumable class shards;
- pointwise, DPP, MMR, k-medoids, and same-objective greedy baselines;
- ADM assembly/evaluation wrappers and DINO/CLIP robustness audits;
- matrix checksums, energy-equivalence checks, int14 Ising exports, and strict physical-return auditing without vote or cardinality repair.

For each formal run we record the model/configuration hash, class and candidate indices, parent lineage, selected set, selector objective, generation archive hash, and evaluator output. All formal image archives contain exactly 50,000 finite uint8 samples with 50 samples per class. The SiT curve reuses only deterministic baseline samples where the selected indices are identical. Each control archive is assembled from its own selected candidate indices.

Submission checks. The independent dynamic $B = 1, 2, 3, 4$ scaling replication, matched Inception-DPP/Moment-greedy comparison, and frozen-pool linear/shuffled interaction ablation are complete. Their principal comparisons are discussed in the main paper; full metric tables are collected in this appendix. For the latter, objective weights, quality floor, swap budget, shuffle namespace, and image gates were frozen before the independent noise namespaces generated their 200K candidates. The independent scaling gate passes, whereas the final-pool 0.995-quality experiment misses its ADM precision non-inferiority gate. Independent full-class SiT retention and DiT repeated-branching confirmations are complete (Appendix E). The calibrated DiT DPP comparison, ordinary CFG 3 population, conditional intervals and CLIP evaluation are complete (Appendix J). The corrected 5,913/5,771-spin physical returns pass the original 99.5% quality floor. Their native-SA 10 ms energy comparison and the separate raw-CIM/long-budget-reference ADM-50K evaluations are complete (Appendix I.1). The historical reference comparison, ordinary Kaiwu SA controls, fixed-primary historical 10 ms ADM evaluations, and six conditional P/R comparisons are also included. The timing protocol above specifies the per-attempt records for further device comparisons.