STRUCTURAL ILLUMINATION UNDER CONTROL–REALIZATION MISMATCH FOR COMPOSITIONAL TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Compositional text-to-image generation is typically optimized to produce a faithful *single* image, yet the same compositional constraint can admit many valid realizations with different spatial configurations. We study the complementary problem of *structural repertoire discovery*: given a fixed generation budget, how can we uncover diverse realizations that satisfy the same compositional constraint? We formalize this objective as *feasible structural illumination*, which seeks to cover distinct structural niches using constraint-satisfying generated images. A key challenge is *control–realization mismatch*: inference-time search selects among candidate control configurations that specify intended structure, but stochastic generation may realize a different structure in the resulting image. To address this mismatch, we introduce Archive-Guided Structural Illumination (AGSI), which uses cheap structural proxies derived from these controls to guide search. Only constraint-satisfying generated images can expand structural coverage, and their realized structure determines the niche they occupy. On GenEval spatial relations with GLIGEN, AGSI achieves higher realized structural coverage than strong search baselines under matched generation budgets. The same principle transfers to exact counting and T2I-CompBench++ prompts, and remains effective when replacing GLIGEN with InstanceDiffusion. A matched ablation further shows that tracking discovered niches from realized outputs rather than proxy predictions improves coverage on both tasks. Notably, AGSI achieves broader coverage than the strongest baseline with a similar or lower fraction of feasible generations, indicating that its gains come from using feasible generations more effectively rather than simply producing more of them.

1 INTRODUCTION

Text-to-image (T2I) models have made rapid progress in visual fidelity and prompt following, yet reliably satisfying *compositional prompts* remains challenging when a prompt specifies multiple objects, attributes, counts, or spatial relations. Recent work improves compositional generation through attention guidance, explicit spatial grounding, layout planning, and instance-level control (Feng et al., 2022; Li et al., 2023; Chefer et al., 2023; Xie et al., 2023; Feng et al., 2023; Zhou et al., 2024; Yang et al., 2024; Binyamin et al., 2025). In parallel, inference-time methods increasingly use additional computation to search, select, or optimize generations according to alignment or preference objectives (Guo et al., 2024; Eyring et al., 2024; Ma et al., 2025). Despite these advances, most inference-time approaches still seek to improve or select a single output for a given prompt.

We study a complementary objective motivated by a simple observation: a compositional prompt can admit multiple structurally distinct realizations that all satisfy the same constraint. For example, "a dog to the left of a chair" remains valid across different object separations and relative scales. Optimizing only for a highest-scoring sample leaves this multiplicity unexplored. We therefore study *structural repertoire discovery*: under a fixed budget of expensive image generations, the goal is to discover a repertoire of structurally distinct, constraint-satisfying realizations rather than a single best output.

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.

We formalize this problem as *feasible structural illumination*. Each feasible generated image is assigned to a discrete structural niche according to its realized structure, and the objective is to maximize the number of distinct feasible niches discovered. This connects naturally to quality-diversity (QD) illumination, which seeks repertoires rather than a single optimum (Chatzilygeroudis et al., 2021; Qin et al., 2026; Mouret & Clune, 2015). Building on flexible conditional generation in diffusion-based T2I models (Rombach et al., 2022; Saharia et al., 2022; Gu et al., 2022), controllable methods provide generation controls such as grounding boxes or instance layouts (Yang et al., 2023; Zheng et al., 2023; Li et al., 2023; Bar-Tal et al., 2023; Zhou et al., 2024), providing a control space over intended structures that can be searched at inference time. However, a control specifies an intended structure, while stochastic generation may realize a different structure in the resulting image. We call this discrepancy *control–realization mismatch*. Consequently, broad exploration of the control space or its proxy niches need not yield broad realized structural coverage.

This motivates our central principle: *control-derived proxy structure guides where to search, while generated outputs determine what has been discovered*. We instantiate this principle as *Archive-Guided Structural Illumination* (AGSI). AGSI maintains realized feasible niche occupancy as its search state. For each candidate control, its intended layout provides a cheap proxy structural niche, and AGSI prioritizes candidates whose proxy niches are absent from the current realized occupancy. After generation, the realized occupancy expands only when the output is feasible and falls in a previously undiscovered realized niche.

We evaluate AGSI under fixed generation budgets, with spatial relations as the primary setting. On the full GenEval spatial-relation setting, AGSI discovers 7.51 of 9 realized niches at $B = 64$, compared with 7.12 for Proxy-Balanced and 7.10 for BOP-Elites. A matched archive-state ablation confirms that using realized rather than proxy occupancy improves structural coverage on both tasks. The same search principle transfers to exact counting, T2I-CompBench++ (Huang et al., 2025), and InstanceDiffusion (Wang et al., 2024). Further diagnostics show that the gains are not explained by higher feasibility alone and are associated with more effective discovery of new realized niches, while a latent-noise optimization comparison under a fixed control sequence suggests that stronger realization does not fully substitute for structural search.

Our contributions are:

- We formulate compositional T2I inference as *structural repertoire discovery* under a fixed generation budget and instantiate it as *feasible structural illumination*, where the objective is to maximize coverage of feasible realized structural niches.
- We identify *control–realization mismatch* as a key challenge in searching a control space for realized structural coverage, and introduce AGSI, which uses cheap control-derived proxy structure to guide search while measured feasibility and realized structure determine archive occupancy.
- We demonstrate improved realized feasible structural coverage under fixed generation budgets across task, benchmark, and generator shifts. A matched archive-state ablation isolates the benefit of realized occupancy, while diagnostics reveal higher new-niche discovery per feasible generation despite similar or lower feasibility.

2 Related Work

2.1 Compositional Text-to-Image Generation and Inference-Time Search

Compositional text-to-image generation requires jointly representing multiple objects, attributes, counts, and relations in a single image. Attention-based methods improve compositional correspondence: Structured Diffusion Guidance incorporates linguistic structure into cross-attention (Feng et al., 2022) and Attend-and-Excite strengthens attention to neglected subject tokens (Chefer et al., 2023). More explicit structural control is provided by GLIGEN (Li et al., 2023), BoxDiff (Xie et al., 2023), and MIGC (Zhou et al., 2024), which exploit grounding boxes or instance-level control. Beyond direct guidance, recent methods use planning or feedback for complex prompts: RPG decomposes generation into region-wise subtasks via multimodal-LLM planning (Yang et al., 2024), while IterComp and SILMM improve compositionality through iterative feedback (Zhang et al., 2025) and self-improvement (Qu et al., 2025).

A complementary line of work improves generation through additional inference-time computation, e.g., by optimizing diffusion latents with image-level objectives (DOODL (Wallace et al., 2023)), refining initial noise via attention-derived signals (InitNO (Guo et al., 2024)) or learned reward models (ReNO (Eyring et al., 2024)). Other methods allocate computation across attempts, such as Best-of-N sampling, textual feedback refinement (Reflect-DiT (Li et al., 2025)), or search over test-time objectives (Zhang et al., 2026). Despite these differences, all such methods improve or select individual outputs. We instead allocate a fixed generation budget to discovering a repertoire of feasible, structurally distinct realizations.

2.2 Quality-Diversity and Stochastic Realization

Quality-Diversity (QD) seeks repertoires of high-performing, behaviorally diverse solutions rather than a single optimum. MAP-Elites partitions behavior space into niches and retains the best solution per niche (Mouret & Clune, 2015). CMA-ME combines this archiving with CMA-ES adaptation (Fontaine et al., 2020), and CMA-MAE further adds annealed per-niche thresholds to balance exploration and optimization (Fontaine & Nikolaidis, 2023). For expensive black-box evaluations, BOP-Elites uses Gaussian-process surrogates for sample-efficient repertoire discovery (Kent et al., 2024), while ECI instead seeks diverse configurations within satisfactory regions (Malkomes et al., 2021). Most related to our work, QDHF learns diversity metrics from human similarity judgments and applies QD to text-to-image generation (Ding et al., 2024). These works motivate our repertoire-level objective of feasible structural illumination.

Another relevant line of QD work distinguishes the object being searched from the behavior revealed through evaluation. In QD-RL, PGA-MAP-Elites combines archive-based search with policy-gradient updates (Nilsson & Cully, 2021), while QD-PG directs policies toward underexplored behavior regions via diversity policy gradients (Pierrot et al., 2022). Uncertain-QD more directly addresses stochastic fitness and behavioral descriptors through repeated evaluation or distributional modeling (Justesen et al., 2019; Flageat & Cully, 2023). Our setting assigns stochastic realization a different role: different feasible images generated from the same control may occupy different structural niches and therefore constitute distinct discoveries, rather than merely repeated observations used to estimate an uncertain descriptor. Building on output-based QD archives, AGSI compares cheap control-derived proxy niches against realized feasible occupancy to target undiscovered structures without generating images for unselected candidates.

3 Problem Formulation: Feasible Structural Illumination

3.1 From Best-Sample Optimization to Structural Repertoires

Let $\mathcal{P}$ denote the prompt space and $\mathcal{Y}$ the image space. For a prompt $p \in \mathcal{P}$ and generation budget $B \in \mathbb{N}_+$, inference-time T2I optimization commonly seeks a single high-scoring image $x^\star \in \arg\max_{x \in \mathcal{X}_B(p)} R_p(x)$, where $\mathcal{X}_B(p) \subseteq \mathcal{Y}$ is the generated image set and $R_p : \mathcal{Y} \to \mathbb{R}$ is a reward function. Yet the same compositional constraint can admit multiple valid realizations with different structures. We instead seek a repertoire of such feasible structures under the same budget.

We consider a stochastic controllable T2I generator G that produces $x = G(p, u, z) \in \mathcal{Y}$ from prompt p, control $u \in \mathcal{U}_p$, and randomness $z \in \mathcal{Z}$. We define a hard feasibility function $F_p : \mathcal{Y} \to \{0, 1\}$, with $F_p(x) = 1$ when x satisfies the compositional constraint specified by prompt p. For each feasible output, we measure a realized structural descriptor $b_p(x)$ and map it to a discrete niche by

$$c_p(x) = C_p(b_p(x)), \tag{1}$$

where C_p denotes the task-dependent discretization mapping and $c_p(x)$ the resulting realized niche, instantiated for each task in Sec. 4.3 and full discretization rules are provided in Appendix B.2.

After t generations, the realized feasible occupancy is

$$\mathcal{O}_t = \{c_p(x_i) \,|\, 1 \le i \le t,\ F_p(x_i) = 1\}. \tag{2}$$

We define *feasible structural illumination* as maximizing the expected coverage of this occupancy under a fixed generation budget:

$$\pi^\star \in \arg\max_{\pi \in \Pi} \mathbb{E}_\pi \left[|\mathcal{O}_B|\right], \tag{3}$$

where Π denotes the class of sequential search strategies for selecting generation controls based on previous observations.

3.2 Control–Realization Mismatch and Its Consequences

The challenge is that each control must be selected before its realized structural niche is known. Before generation, however, a control u provides a cheap indication of its intended structure. We denote the corresponding control-derived proxy descriptor by $\tilde{b}_p(u)$ and map it to a proxy structural niche using the same niche mapping C_p applied to the realized structural descriptor $b_p(x)$:

$$\tilde{c}_p(u) = C_p\Big(\tilde{b}_p(u)\Big)\,. \tag{4}$$

Thus, $\tilde{b}_p(u)$ and $\tilde{c}_p(u)$ are available before generation, whereas $b_p(x)$ and $c_p(x)$ are measured from the generated image.

Because generation is stochastic, the intended and realized structures need not coincide:

$$\tilde{c}_p(u) \neq c_p(G(p,u,z))\,. \tag{5}$$

We call this discrepancy *control–realization mismatch*. Consequently, broad exploration of controls or proxy niches need not yield broad realized coverage, since a control targeting an uncovered proxy niche may still realize an already occupied niche. Maximizing feasibility alone may likewise produce structurally redundant outputs.

Therefore, we treat proxy structure as informative for where to search, but not as evidence of what has been discovered. This motivates our method: control-derived proxy structure guides search, while discovery state is updated only from feasible realized outputs.

4 Archive-Guided Structural Illumination

4.1 Realized Feasible Archive

At generation step t, the realized feasible archive comprises the occupancy $\mathcal{O}_t$, which serves as the archive state for subsequent search, and per-niche histories of feasible outputs $\{\mathcal{H}_{j,t}\}_{j\in\mathcal{O}_t}$, retained for downstream representative selection. After generating x_t, $\mathcal{O}_t$ is updated as

$$\mathcal{O}_t = \begin{cases} \mathcal{O}_{t-1} \cup \{c_p(x_t)\}, & F_p(x_t) = 1, \\ \mathcal{O}_{t-1}, & \text{otherwise.} \end{cases} \tag{6}$$

Thus, only the realized structure of a feasible generated image can expand $\mathcal{O}_t$; the proxy niche of its control does not.

For each occupied niche $j \in \mathcal{O}_t$, its history contains all feasible outputs assigned to that niche:

$$\mathcal{H}_{j,t} = \{x_i \mid 1 \le i \le t,\ F_p(x_i) = 1,\ c_p(x_i) = j\}\,. \tag{7}$$

After search, a representative can be selected from $\mathcal{H}_{j,B}$ using an application-specific utility $r(x)$, i.e., $x_j^\star(r) = \arg\max_{x\in\mathcal{H}_{j,B}} r(x)$. This post-hoc choice does not affect $\mathcal{O}_t$ or subsequent search.

4.2 Archive-Guided Search

Figure 1 summarizes the search procedure. For a candidate control u, the expected one-step increase in realized feasible coverage is $\Delta_t(u) = \Pr_z[F_p(x) = 1,\ c_p(x) \notin \mathcal{O}_{t-1}]$, where $x = G(p,u,z)$. Estimating this quantity through repeated generation for each candidate would consume additional generation budget. Because the proxy is informative but imperfect, we instead combine the candidate's proxy niche with the current realized occupancy to define a binary search priority:

$$a_t(u) = \mathbf{1}[\tilde{c}_p(u) \notin \mathcal{O}_{t-1}]\,, \tag{8}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function and undefined proxy niches receive priority zero. Thus, $a_t(u)$ is a heuristic search priority, not an estimate of $\Delta_t(u)$.

We first use B_0 unguided generations from the same predefined space of valid controls to initialize $\mathcal{O}_{B_0}$. For each $t > B_0$, we sample K valid candidate controls and evaluate $a_t(u)$ for each. If any candidate has priority one, we select the first such candidate in proposal order; otherwise, we uniformly sample from the same pool. Only the selected control is generated, and its output updates $\mathcal{O}_t$ according to Eq. 6. The complete procedure is given in Appendix A.

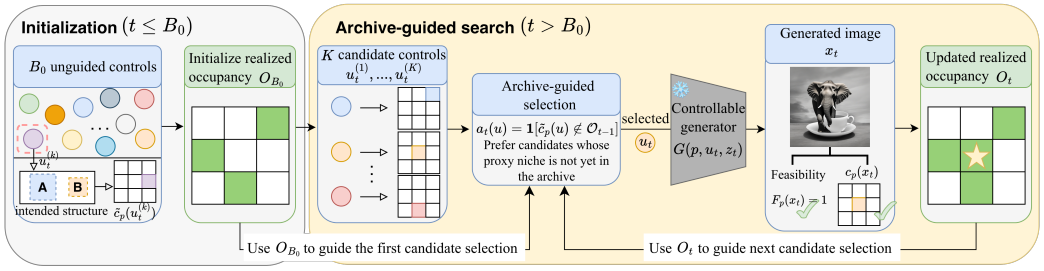

Figure 1: **Overview of Archive-Guided Structural Illumination.** After B_0 unguided generations initialize realized feasible occupancy, AGSI selects among K candidate controls using proxy niches and current occupancy $\mathcal{O}_t$, then updates the archive from measured feasibility and realized structure.

4.3 Task Instantiations

The search rule in Eq. 8 is shared across tasks, while the control u, structural descriptor b_p, feasibility criterion F_p, and niche mapping C_p are task-specific. In both tasks, the proxy descriptor $\tilde{b}_p(u)$ is computed using the same descriptor definition as $b_p(x)$, but from the intended control layout rather than the realized image structure. Proxy and realized descriptors are mapped by the same task-specific C_p into a 3×3 archive. The task-specific discretization rules, are provided in Appendix B.2.

Spatial relations. For a prompt specifying a spatial relation between objects A and B, the control u specifies their intended grounding layout. From the detected object geometry, we define

$$b_{\text{spatial}}(x) = \left(s_{\parallel}(x), \log \frac{a_A}{a_B} \right), \tag{9}$$

where $s_{\parallel}(x)$ measures the separation along the requested relation axis, with its sign normalized so that larger positive values indicate greater separation, and a_A and a_B denote the detected object areas. We set $F_p(x) = 1$ when both objects are present and the requested spatial relation is satisfied.

Exact counting. For a target count n, the control u specifies the intended instance layout. For detected instance centers $(x_i, y_i)_{i=1}^{n}$, we define

$$b_{\text{count}}(x) = (\Delta_x, \Delta_y), \qquad \Delta_x = \max_i x_i - \min_i x_i, \quad \Delta_y = \max_i y_i - \min_i y_i. \tag{10}$$

This descriptor captures the horizontal and vertical extent of the realized group and is invariant to instance ordering. We set $F_p(x) = 1$ when exactly n instances of the requested object are detected.

5 Experiments

5.1 Experimental Setup

Tasks and benchmarks. We use spatial relations as the primary setting and exact counting as a cross-task transfer setting. On GenEval (Ghosh et al., 2023), we evaluate all 100 spatial-relation prompts (left, right, above, below) and all 80 exact-counting prompts ($n \in \{2, 3, 4\}$). For cross-benchmark transfer, we use all T2I-CompBench++ prompts (Huang et al., 2025) that fall within these predefined task settings, yielding 136 spatial-relation and 47 exact-counting prompts. Both tasks use the 3×3 structural archives defined in Sec. 4.3. Hard feasibility is measured with the GenEval Mask2Former Swin-S evaluator: spatial prompts require both objects and the requested relation, while counting requires exactly n detected instances. Full discretization and evaluator settings are provided in Appendix B.2.

Generation and search configuration. Unless otherwise stated, we use Stable Diffusion v1.5 with GLIGEN box grounding at 512×512 resolution (Li et al., 2023), with InstanceDiffusion (Wang

Table 1: **Structural illumination on Full GenEval.** We report realized feasible coverage at fixed generation budgets and AUC/B over the search trajectory. Best and second-best values are shown in bold and underlined, respectively.

	Spatial relations (100 prompts)				Exact counting (80 prompts)		
Method	$\mathrm{Cov}_{16}\uparrow$	$\mathrm{Cov}_{32}\uparrow$	$\mathrm{Cov}_{64}\uparrow$	AUC/$B\uparrow$	$\mathrm{Cov}_{16}\uparrow$	$\mathrm{Cov}_{32}\uparrow$	AUC/$B\uparrow$
Random	4.017	5.087	6.070	4.657	3.521	5.096	3.315
Sobol	4.003	5.203	6.150	4.739	3.704	5.358	3.475
Structural-ME	3.583	4.557	5.353	4.182	3.663	5.192	3.380
CMA-MAE	3.947	4.983	5.813	4.544	3.625	5.246	3.379
BOP-Elites	3.987	<u>6.043</u>	7.097	5.330	3.521	<u>5.633</u>	3.489
ECI	3.723	4.693	5.410	4.275	3.154	4.354	2.939
Proxy-Balanced	<u>4.483</u>	6.010	<u>7.117</u>	<u>5.388</u>	<u>3.975</u>	5.454	<u>3.591</u>
AGSI	**4.877**	**6.410**	**7.507**	**5.718**	**4.238**	**6.258**	**3.923**

et al., 2024) used for generator transfer. All methods search the same predefined valid control space for each task; its parameterization, numerical bounds, and generator settings are given in Appendix B.1. We use three search seeds, with $B = 64$ for the primary spatial setting and $B = 32$ for the cross-task, cross-benchmark, and cross-generator transfer settings. AGSI uses $B_0 = 8$ unguided generations and $K = 32$ candidate controls per subsequent step; proxy evaluations are budget-free. Descriptors, discretizations, and search hyperparameters are fixed before full-benchmark evaluation.

Baselines and metrics. On GenEval, we compare against Random, Sobol (Sobol, 1967), Structural-ME (Mouret & Clune, 2015), CMA-MAE (Fontaine & Nikolaidis, 2023), BOP-Elites (Kent et al., 2024), ECI (Malkomes et al., 2021), and Proxy-Balanced, which balances selections across proxy niches without using feasibility or realized occupancy. Cross-benchmark and cross-generator transfer experiments use Random, Sobol, and BOP-Elites. All initialization generations count toward the same budget B; baseline details are in Appendix B.3. Our primary metric is realized feasible structural coverage, $\mathrm{Cov}_B = |\mathcal{O}_B|$; intermediate-budget coverage and normalized AUC (AUC/B) summarize search progress. Results are averaged over three seeds within each prompt, with paired bootstrap confidence intervals computed over prompts (Koehn, 2004).

5.2 Does Archive-Guided Search Improve Structural Illumination?

We first evaluate whether using realized archive occupancy to guide search improves structural illumination under a fixed generation budget. As shown in Table 1, AGSI achieves the highest realized feasible coverage at every reported budget in the spatial-relations setting. At $B = 64$, it discovers 7.507 of the 9 structural niches on average, compared with 7.117 for Proxy-Balanced, 7.097 for BOP-Elites, and $6.070/6.150$ for Random/Sobol. The prompt-paired gains at $B = 64$ are $+0.390$ niches over Proxy-Balanced (95% CI $[0.277, 0.507]$) and $+0.410$ over BOP-Elites (95% CI $[0.227, 0.590]$). Per-niche decomposition in Figure 2 further shows positive occupancy differences in eight of the nine niches relative to both comparators, indicating that the gain is distributed across the realized archive rather than driven by a single niche. AGSI also achieves the highest AUC/B (5.718 vs. 5.388 for Proxy-Balanced and 5.330 for BOP-Elites), indicating more efficient niche discovery over the search trajectory rather than an advantage confined to the final budget.

To complement this aggregate analysis with an instance-level view, Figure 3 visualizes the realized structural repertoires discovered by the three methods for the same prompt, search seed, and generation budget. BOP-Elites discovers $6/9$ realized niches, Proxy-Balanced $7/9$, and AGSI $8/9$. The additional occupied cells illustrate how the aggregate coverage gains translate into a broader repertoire of feasible realized structures, rather than additional samples within already occupied niches.

5.3 Does Realized Occupancy Matter for Search?

For AGSI on GenEval, proxy and realized niches differ in 36.5% of feasible generations for spatial relations ($B = 64$) and 25.7% for exact counting ($B = 32$), pooling feasible generations across all prompts and three seeds. This mismatch indicates that proxy occupancy may misrepresent realized

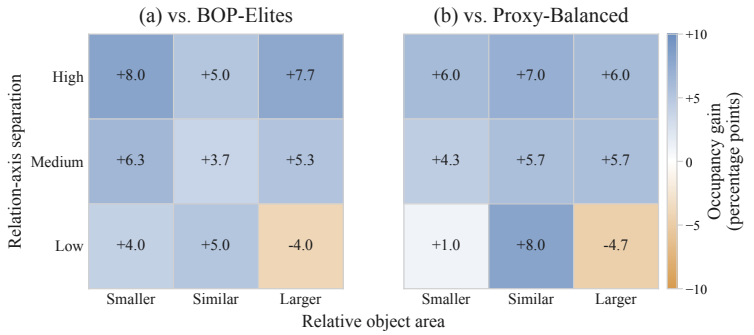

Figure 2: **Per-niche occupancy gains on Full GenEval spatial relations at $B = 64$.** Cells report AGSI minus baseline final realized-niche occupancy rates in percentage points, aggregated over 100 prompts and three seeds.

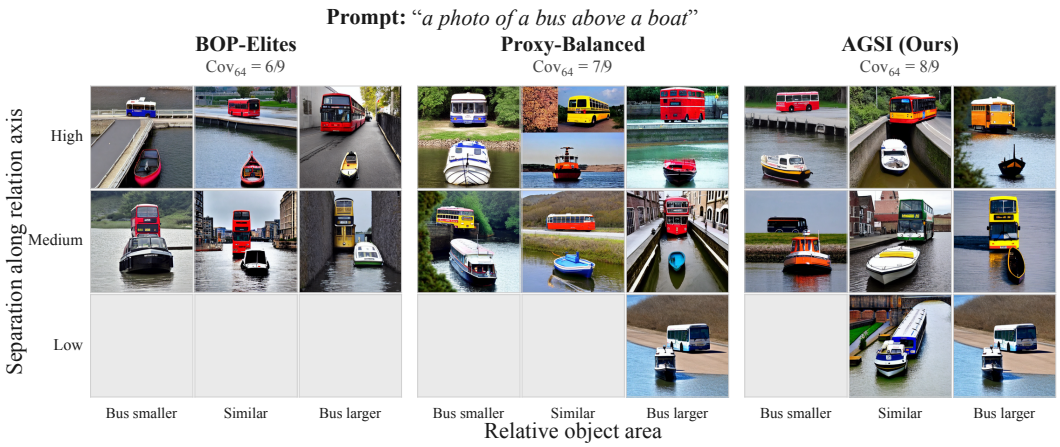

Figure 3: **Realized structural repertoires on an illustrative Full GenEval spatial-relations example at $B = 64$.** Each occupied cell shows a feasible representative selected using the same frozen post-hoc rule across methods; gray cells denote undiscovered realized niches.

coverage. We therefore use a matched archive-state ablation to test whether this discrepancy affects subsequent search.

To isolate the effect of realized feedback, we compare the standard *Realized-Archive* variant of AGSI with *Proxy-Archive* variant. Both update search occupancy only after feasible generations, differing only in whether they insert realized output niches or control-derived proxy niches. As shown in Table 2, Realized-Archive improves final coverage by $+0.873$ niches on spatial relations (95% CI $[0.75, 1.00]$) and $+0.525$ on exact counting (95% CI $[0.41, 0.64]$), with higher AUC/B on both tasks. Figure 4 further shows that the trajectories coincide during the shared $B_0 = 8$ initialization and diverge once archive-guided search begins, with the gap generally increasing thereafter, indicating that realized occupancy affects subsequent search rather than merely measuring discovery.

5.4 Does the Principle Generalize Across Tasks, Benchmarks, and Generators?

We next test whether the archive-guided principle transfers beyond the primary spatial setting. Across these experiments, we keep the search procedure fixed—including $B_0 = 8$, $K = 32$, and the same realized-occupancy update rule—while varying one axis at a time: task formulation, benchmark prompt distribution, or controllable generator.

Table 2: **Matched archive-state ablation.** Realized-Archive and Proxy-Archive differ only in the occupancy state used to guide search and are evaluated using realized feasible coverage. Δ denotes the prompt-paired Realized-Archive minus Proxy-Archive difference.

	Spatial relations (100 prompts)				Exact counting (80 prompts)		
Search state	$\mathrm{Cov}_{16} \uparrow$	$\mathrm{Cov}_{32} \uparrow$	$\mathrm{Cov}_{64} \uparrow$	$\mathrm{AUC}/B \uparrow$	$\mathrm{Cov}_{16} \uparrow$	$\mathrm{Cov}_{32} \uparrow$	$\mathrm{AUC}/B \uparrow$
Proxy-Archive	4.490	5.730	6.633	5.143	4.104	5.733	3.733
Realized-Archive	**4.877**	**6.410**	**7.507**	**5.718**	**4.238**	**6.258**	**3.923**
Δ (Realized − Proxy)	+0.387	+0.680	+0.873	+0.575	+0.133	+0.525	+0.190

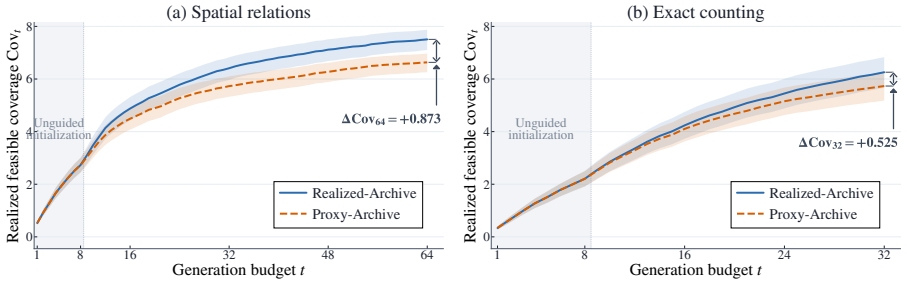

Figure 4: **Search dynamics for the matched archive-state ablation.** Curves show mean realized feasible coverage Cov_t with 95% prompt-bootstrap confidence intervals. The first $B_0 = 8$ generations are shared unguided initialization.

Across tasks. Replacing spatial relations with exact counting changes both the structural descriptor and feasibility criterion while retaining the same search rule. As shown in Table 1, at $B = 32$ AGSI exceeds BOP-Elites by $+0.625$ niches (95% CI $[0.396, 0.858]$), with positive gains for $n = 2, 3, 4$ and all three 95% confidence intervals excluding zero.

Across benchmark prompts. As shown in Table 3, on the spatial-relation and exact-counting subsets of T2I-CompBench++ (Huang et al., 2025), AGSI exceeds BOP-Elites at $B = 32$ by $+0.373$ niches on Spatial (95% CI $[0.20, 0.54]$) and $+0.319$ on Count (95% CI $[0.06, 0.60]$). These gains persist under a benchmark prompt-distribution shift without benchmark-specific retuning.

Across generators. As shown in Table 3, replacing GLIGEN with InstanceDiffusion yields gains over BOP-Elites at $B = 32$ of $+1.120$ niches on Spatial (95% CI $[0.90, 1.34]$) and $+0.512$ on Count (95% CI $[0.28, 0.76]$), indicating that the advantage is not tied to the GLIGEN-specific mapping from controls to realized image structures.

Taken together, these results support transfer of the same archive-guided principle across changes in structural formulation, prompt distribution, and controllable generator.

5.5 Where Does the Coverage Gain Come From?

We first ask whether balancing intended structures alone is sufficient to explain the coverage gain. Although proxy balancing is beneficial, AGSI still improves coverage at $B = 32$ by $+0.400$ niches on Spatial (95% CI $[0.260, 0.540]$) and $+0.804$ on Count (95% CI $[0.588, 1.021]$). The gain can also not be explained by higher feasibility: relative to Proxy-Balanced, AGSI has nearly unchanged feasibility on Spatial ($\Delta = -0.003$, 95% CI $[-0.011, 0.005]$) and 4.1 percentage points (95% CI $[-5.7, -2.5]$) lower feasibility on Count. Among feasible generations, however, the probability of opening a new realized niche, $P(\text{new} \mid F = 1)$, is higher by $+3.6$ and $+15.1$ percentage points, respectively. This is consistent with the feasible structural illumination objective in Eq. 3, which rewards expansion of realized occupancy rather than feasibility alone. Full diagnostics and the matched-niche preference audit are provided in Appendix D.1 and Appendix C.

Table 3: **Generalization across benchmarks and generators.** We report realized feasible coverage and AUC/B under benchmark- and generator-transfer settings. Best and second-best values are shown in bold and underlined, respectively.

	Spatial relations			Exact counting		
Method	$\mathrm{Cov}_{16}\uparrow$	$\mathrm{Cov}_{32}\uparrow$	AUC/$B\uparrow$	$\mathrm{Cov}_{16}\uparrow$	$\mathrm{Cov}_{32}\uparrow$	AUC/$B\uparrow$
T2I-CompBench++ (Spatial: 136 prompts; Count: 47 prompts)						
Random	5.250	6.328	4.798	4.199	5.936	3.922
Sobol	<u>5.502</u>	6.407	4.960	<u>4.284</u>	6.028	3.940
BOP-Elites	5.250	<u>7.806</u>	<u>5.324</u>	4.199	<u>6.298</u>	<u>4.101</u>
AGSI	**6.816**	**8.179**	**5.955**	**4.929**	**6.617**	**4.416**
InstanceDiffusion on GenEval (Spatial: 100 prompts; Count: 80 prompts)						
Random	4.833	5.937	4.431	4.317	6.058	3.992
Sobol	<u>4.993</u>	6.070	4.586	4.263	5.992	3.946
BOP-Elites	4.833	<u>6.793</u>	<u>4.753</u>	<u>4.321</u>	<u>6.463</u>	<u>4.145</u>
AGSI	**6.447**	**7.913**	**5.636**	**5.096**	**6.975**	**4.577**

Finally, we ask whether stronger realization of a fixed control sequence can substitute for structural search. In the matched Spatial comparison at $B = 32$, Noise Diffusion (Miao et al., 2025) optimizes only the initial noise under the frozen Sobol controls, increasing coverage from 5.21 to 5.88 and feasibility from 0.512 to 0.697. Yet the matched AGSI run reaches 6.36 niches and exceeds Noise Diffusion by 0.48 cells (95% CI $[0.14, 0.80]$), despite lower feasibility (0.487). Noise Diffusion uses 2580 versus 30 UNet evaluations per returned image, so the comparison is generation-budget matched rather than compute matched. Full details are provided in Appendix D.2. Together, these results indicate that proxy balancing and latent optimization both help, but neither substitutes for allocating generations according to what has actually been discovered.

6 Discussion and Conclusion

We studied compositional T2I inference from a repertoire-level perspective: rather than seeking a single preferred output, *feasible structural illumination* aims to discover structurally distinct, constraint-satisfying realizations under a fixed generation budget. The central challenge is *control–realization mismatch*: a control specifies an intended structure, but stochastic generation may realize a different one. AGSI addresses this mismatch with a simple principle—control-derived proxy structure guides search, while measured feasibility and realized structure determine archive occupancy. Experiments across task, benchmark, and generator shifts, together with matched archive-state ablations and diagnostics, support the benefit of realized-occupancy feedback and show that coverage gains need not come from higher feasibility. Together, these results suggest that allocating generation budget according to what has actually been discovered is complementary to both balancing intended structures and improving realization under fixed controls.

Limitations and future work. Our current instantiations rely on task-specific, low-dimensional structural descriptors and external image evaluators, so they capture only selected structural variation and inherit evaluator errors. AGSI further assumes that the control-derived proxy is sufficiently informative to guide search. Its objective targets structural discovery rather than within-niche perceptual quality, so broader coverage does not guarantee uniformly high-quality representatives. Finally, our evaluation matches methods by generation budget rather than total inference compute or latency. Future work could explore richer or learned structural representations and proxies, more robust realized-structure evaluation, joint structural-coverage and quality optimization, and compute-aware evaluation across broader compositional tasks and controllable generators.

AI use statement

In this work, we used generative AI tools to aid and polish the writing, and for retrieval and discovery tasks, such as identifying and summarizing potentially relevant related work. We have not used

generative AI tools for research ideation or execution, drafting sections of the paper, generating synthetic datasets, or proving mathematical claims; these uses are not applicable to this work.

Additionally, we used generative AI tools for tasks with recommended disclosure, such as paraphrasing, grammar checking, and minor stylistic refinements of the text. We have reviewed all AI-assisted work: all AI-suggested citations and related-work leads were verified by the authors through a manual literature survey, and all AI-polished text was carefully read, fact-checked, and revised by the authors to ensure technical accuracy and consistency with our results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13242–13251. IEEE, 2025.

Konstantinos Chatzilygeroudis, Antoine Cully, Vassilis Vassiliades, and Jean-Baptiste Mouret. Quality-diversity optimization: a novel branch of stochastic optimization. In *Black box optimization, machine learning, and no-free lunch theorems*, pp. 109–135. Springer, 2021.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Li Ding, Jenny Zhang, Jeff Clune, Lee Spector, and Joel Lehman. Quality diversity through human feedback: towards open-ended diversity-driven optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in neural information processing systems*, 36: 18225–18250, 2023.

Manon Flageat and Antoine Cully. Uncertain quality-diversity: evaluation methodology and new methods for quality-diversity in uncertain domains. *IEEE Transactions on Evolutionary Computation*, 28(4):891–902, 2023.

Matthew Fontaine and Stefanos Nikolaidis. Covariance matrix adaptation map-annealing. In *Proceedings of the genetic and evolutionary computation conference*, pp. 456–465, 2023.

Matthew C Fontaine, Julian Togelius, Stefanos Nikolaidis, and Amy K Hoover. Covariance matrix adaptation for the rapid illumination of behavior space. In *Proceedings of the 2020 genetic and evolutionary computation conference*, pp. 94–102, 2020.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10686–10696. IEEE, 2022.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9380–9389. IEEE, 2024.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

Niels Justesen, Sebastian Risi, and Jean-Baptiste Mouret. Map-elites for noisy domains by adaptive sampling. In *Proceedings of the genetic and evolutionary computation conference companion*, pp. 121–122, 2019.

Paul Kent, Adam Gaier, Jean-Baptiste Mouret, and Juergen Branke. Bayesian optimization for quality diversity search with coupled descriptor functions. *IEEE Transactions on Evolutionary Computation*, 29(2):302–316, 2024.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pp. 388–395, 2004.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Arsh Koneru, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Reflect-dit: Inference-time scaling for text-to-image diffusion transformers via in-context reflection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15657–15668. IEEE, 2025.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Scaling inference time compute for diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2523–2534. IEEE, 2025.

Gustavo Malkomes, Bolong Cheng, Eric H Lee, and Mike Mccourt. Beyond the pareto efficient frontier: Constraint active search for multiobjective experimental design. In *International Conference on Machine Learning*, pp. 7423–7434. PMLR, 2021.

Boming Miao, Chunxiao Li, Xiaoxiao Wang, Andi Zhang, Rui Sun, Zizhe Wang, and Yao Zhu. Noise diffusion for enhancing semantic faithfulness in text-to-image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23575–23584. IEEE, 2025.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

Olle Nilsson and Antoine Cully. Policy gradient assisted map-elites. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 866–875, 2021.

Thomas Pierrot, Valentin Macé, Felix Chalumeau, Arthur Flajolet, Geoffrey Cideron, Karim Beguir, Antoine Cully, Olivier Sigaud, and Nicolas Perrin-Gilbert. Diversity policy gradient for sample efficient quality-diversity optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1075–1083, 2022.

Haoxiang Qin, Yi Xiang, Hainan Zhang, Yuyan Han, Yuting Wang, Xinrui Tao, and Yiping Liu. A survey on quality-diversity optimization: Approaches, applications, and challenges. *Swarm and Evolutionary Computation*, 100:102240, 2026.

Leigang Qu, Haochuan Li, Wenjie Wang, Xiang Liu, Juncheng Li, Liqiang Nie, and Tat-Seng Chua. Silmm: Self-improving large multimodal models for compositional text-to-image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18497–18508. IEEE, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Ilya M Sobol. Distribution of points in a cube and approximate evaluation of integrals. *USSR Computational mathematics and mathematical physics*, 7:86–112, 1967.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7246–7256. IEEE, 2023.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6232–6242. IEEE, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7418–7427. IEEE, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024.

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14246–14255. IEEE, 2023.

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*, volume 2026, pp. 39133–39166, 2026.

Xinchen Zhang, Ling Yang, Guohao Li, Yaqi Cai, Yong Tang, Yujiu Yang, Mengdi Wang, Bin CUI, et al. Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. In *International Conference on Learning Representations*, volume 2025, pp. 31968–31988, 2025.

Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22490–22499. IEEE, 2023.

Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6818–6828. IEEE, 2024.

Algorithm 1 Archive-Guided Structural Illumination (AGSI)

Require: Prompt p; generator G; valid control space $\mathcal{U}_p$; feasibility function F_p; proxy- and realized-niche mappings $\tilde{c}_p, c_p$; generation budget B; initialization budget B_0; candidate-pool size $K \geq 1$

$\mathcal{O} \leftarrow \emptyset$; $\mathcal{H}_j \leftarrow \emptyset$ for every structural niche j
for $t = 1, \ldots, B$ **do**
  if $t \leq B_0$ **then**
    Sample $u_t \in \mathcal{U}_p$ using the fixed initialization rule
  else
    Sample an ordered pool $(u_t^{(1)}, \ldots, u_t^{(K)})$ of valid controls
    $\mathcal{I}_t \leftarrow \left\{k : \tilde{c}_p(u_t^{(k)}) \text{ is defined and } \tilde{c}_p(u_t^{(k)}) \notin \mathcal{O}\right\}$
    if $\mathcal{I}_t \neq \emptyset$ **then**
      $k_t \leftarrow \min \mathcal{I}_t$ ▷ First priority-one candidate
    else
      Sample k_t uniformly from $\{1, \ldots, K\}$
    end if
    $u_t \leftarrow u_t^{(k_t)}$
  end if
  Sample generator randomness z_t and generate $x_t = G(p, u_t, z_t)$
  if $F_p(x_t) = 1$ **then**
    $j \leftarrow c_p(x_t)$
    $\mathcal{O} \leftarrow \mathcal{O} \cup \{j\}$
    $\mathcal{H}_j \leftarrow \mathcal{H}_j \cup \{x_t\}$
  end if
end for
return $\mathcal{O}$ and $\{\mathcal{H}_j\}_{j \in \mathcal{O}}$

A Algorithmic Details

Algorithm 1 summarizes AGSI. We use $\tilde{c}_p(u) = C_p(\tilde{b}_p(u))$ for the control-derived proxy niche and $c_p(x) = C_p(b_p(x))$ for the realized niche. In the pseudocode, $\mathcal{O}$ and $\mathcal{H}_j$ denote mutable states; after iteration t, they correspond to $\mathcal{O}_t$ and $\mathcal{H}_{j,t}$ defined in Sec. 4.1. Initialization controls and candidate pools follow the fixed valid-control proposal described in Appendix B.1.

The index set $\mathcal{I}_t$ contains exactly the candidates with search priority $a_t(u) = 1$; candidates with undefined proxy niches therefore have priority zero. Only the selected control is passed to the image generator, so each iteration consumes exactly one image generation and the total generation budget is B. Computing proxy niches for the K candidates does not consume this budget.

For an optional representative utility r, a representative is selected after search as $x_j^\star(r) \in \arg\max_{x \in \mathcal{H}_{j,B}} r(x)$ for each $j \in \mathcal{O}_B$. This post-hoc selection does not affect occupancy or the search trajectory.

B Additional Experimental Details

This section provides the experimental details omitted from Sec. 5.1. Within each experiment, all compared methods use the same generator, valid control-input space, structural archive, feasibility evaluator, and expensive-generation budget.

B.1 Generation and Control Inputs

Generator settings. Our primary experiments use Stable Diffusion v1.5 with GLIGEN box grounding (Li et al., 2023). Cross-generator experiments use InstanceDiffusion (Wang et al., 2024). Both generators use 512×512 resolution, PNDM sampling, and FP16 inference. The principal inference settings are summarized in Table 4.

Table 4: Generator settings used in the experiments.

Generator	Steps	CFG	Control setting
GLIGEN	30	7.5	$\beta = 1.0$
InstanceDiffusion	50	7.5	$\alpha = 0.8$, MIS $= 0.36$

Table 5: Frozen spatial control bounds in normalized image coordinates.

Variable	Lower	Upper
m_x, m_y	0.45	0.55
$d_\parallel$	0.24	0.64
$d_\perp$	-0.24	0.24
w_A, h_A, w_B, h_B	0.20	0.50

Spatial relations. We parameterize the control as $u = (m_x, m_y, d_\parallel, d_\perp, w_A, h_A, w_B, h_B)$, where $\mathbf{m} = (m_x, m_y)$ is the midpoint of the intended object centers, $d_\parallel$ and $d_\perp$ specify displacement along and orthogonal to the requested relation axis, and (w_A, h_A) and (w_B, h_B) are the two box sizes. For canonical relation direction $\mathbf{e}_p$,

$$\boldsymbol{\delta} = d_\parallel \mathbf{e}_p + d_\perp \mathbf{e}_p^\perp, \qquad \mathbf{c}_A = \mathbf{m} - \frac{\boldsymbol{\delta}}{2}, \qquad \mathbf{c}_B = \mathbf{m} + \frac{\boldsymbol{\delta}}{2}. \tag{11}$$

with $d_\parallel > 0$ aligned with the requested ordering. The frozen control ranges are given in Table 5. Controls are valid only when both decoded boxes lie within $[0, 1]^2$ and satisfy the requested ordering; overlap is allowed, with no additional separation constraint beyond Table 5.

Exact counting. For target count n, search operates on a normalized control $\xi \in [0, 1]^{2n+2}$, which is decoded into the generation control $u = (c_{x,1}, c_{y,1}, \ldots, c_{x,n}, c_{y,n}, w, h)$, consisting of n instance centers and a shared box size. The decoding is

$$\begin{aligned} w &= 0.20 + 0.30\xi_{2n+1}, \qquad h = 0.20 + 0.30\xi_{2n+2}, \\ c_{x,i} &= \frac{w}{2} + \xi_{2i-1}(1 - w), \quad c_{y,i} = \frac{h}{2} + \xi_{2i}(1 - h), \quad i = 1, \ldots, n. \end{aligned} \tag{12}$$

This mapping keeps all boxes within $[0, 1]^2$; overlap is allowed and no minimum center separation is imposed.

Control proposal. For AGSI, initialization controls and subsequent candidate pools are drawn from the same fixed task-specific proposal. For spatial relations, the eight control variables are sampled independently and uniformly within the bounds of Table 5, with invalid layouts rejected and resampled. For exact counting, controls are sampled uniformly from $[0, 1]^{2n+2}$ and decoded as described below.

B.2 Archive Discretization and Feasibility

Both tasks use 3×3 structural archives. For each task, the same frozen mapping C_p is applied to the control-derived proxy descriptor and to the descriptor measured from the generated image.

Spatial relations. For the spatial descriptor $b_{\text{spatial}}(x) = (s_\parallel(x), \log(a_A/a_B))$, we use cut points 0.25 and 0.45 for relation-axis separation and -0.4 and 0.4 for log-area ratio, yielding nine near/medium/far $\times$ smaller/similar/larger niches.

Exact counting. For the counting descriptor $b_{\text{count}}(x) = (\Delta_x, \Delta_y)$, we use a separate discretization for each target count. For each $n \in \{2, 3, 4\}$, we sample 100,000 valid controls and use the empirical $1/3$ and $2/3$ quantiles of the intended horizontal and vertical extents as bin boundaries, frozen before full-benchmark evaluation. The resulting count-specific boundaries are reported in Table 6. Count-specific quantiles account for the different extent distributions induced by different target counts while applying the same discretization procedure to all methods.

Table 6: **Frozen count-specific archive boundaries.** Boundaries were obtained from empirical tertiles of 100,000 valid controls per target count.

n	Δ_x boundaries	Δ_y boundaries
2	(0.117105, 0.269905)	(0.116940, 0.271779)
3	(0.245788, 0.392952)	(0.244593, 0.392588)
4	(0.325747, 0.455880)	(0.325268, 0.455492)

Table 7: **Principal search settings.** Initialization generations are included in the total expensive-generation budget B.

Method	Initialization	Main setting
Random	–	uniform sampling
Sobol	–	scrambled Sobol sequence
Structural-ME	8 Random generations	Gaussian mutation, $\sigma = 0.20$
CMA-MAE	8 Random generations	$\lambda = 8$, $\mu = 4$, $\sigma_0 = 0.5$, $\alpha = 0.01$
BOP-Elites	16 Random generations	4096 candidates, 8 local restarts
ECI	8 Random generations	$r = 0.5$, 128 QMC samples, 4096 candidates
Proxy-Balanced	$B_0 = 8$ Random generations	proxy-niche balancing without realized occupancy
AGSI	$B_0 = 8$ Random generations	$K = 32$ proxy-evaluated candidates

Hard feasibility. Generated images are evaluated using the GenEval Mask2Former Swin-S detector (Ghosh et al., 2023). Spatial relations use detection confidence threshold 0.3 and relation-position threshold 0.1; feasibility requires both requested objects and the requested relation. Exact counting uses detection threshold 0.9 with NMS disabled and requires exactly n detections of the requested category. The spatial thresholds follow the GenEval evaluation setting, while the higher detection threshold for exact counting provides a stricter criterion for a task that is particularly sensitive to false-positive detections. Realized descriptors are computed from the same detections used for feasibility evaluation.

B.3 Baseline Settings

Table 7 summarizes the main search settings used for AGSI and the GenEval baselines.

AGSI, Proxy-Balanced, Structural-ME, CMA-MAE, and ECI share the same first eight controls from the Random initialization stream, while BOP-Elites uses the first 16. Sobol follows its own sequence from the first evaluation. All initialization generations count toward the total expensive-generation budget B, whereas AGSI's $K = 32$ proxy candidate evaluations require no image generation and therefore do not consume this budget.

All methods are evaluated using the same hard-feasibility and realized-niche definitions. Archive-based baselines use realized structural descriptors for archive assignment, and infeasible outputs do not occupy the reported realized-feasible archive. After initialization, Proxy-Balanced uses AGSI's candidate proposal ($K = 32$), selecting the least-selected proxy niche present in the pool, breaking ties by fixed niche order, and choosing its first proposed control. Selection counts include initialization and increase upon selection regardless of feasibility; realized occupancy is not used.

Soft task-confidence. For baselines requiring a quality signal, we use Qwen2-VL-7B-Instruct to compute

$$s_p(x) = P_{\text{Qwen2-VL}}\left(\text{Yes} \mid x, q_p\right), \tag{13}$$

where q_p is a prompt-specific yes/no query asking whether the requested spatial relation holds or whether the image contains exactly the requested number of instances. The score is the softmax probability of `Yes` against `No` at the first generated token. CMA-MAE uses $s_p(x)$ as fitness and BOP-Elites as its quality objective, while Structural-ME and ECI use it only for within-niche elite replacement. AGSI does not use this score for search.

Table 8: **Post-hoc representative preference comparison between AGSI and BOP-Elites on matched realized niches.** Positive differences favor AGSI; brackets denote 95% prompt-level paired-bootstrap confidence intervals.

Task	Δ HPSv2	Δ ImageReward
Spatial relations	$+0.0001$ $[-0.0005, 0.0007]$	-0.020 $[-0.073, 0.031]$
Exact counting	-0.0005 $[-0.0013, 0.0002]$	-0.019 $[-0.052, 0.015]$

B.4 Statistical Details

In addition to coverage at fixed budgets, we summarize the search trajectory using normalized area under the coverage–budget curve,

$$\mathrm{AUC}/B = \frac{1}{B}\sum_{t=1}^{B} \mathrm{Cov}_t. \tag{14}$$

The feasibility rate is the fraction of generated images satisfying the corresponding hard-feasibility criterion:

$$\mathrm{FeasRate}_B = \frac{1}{B}\sum_{t=1}^{B} F_p(x_t). \tag{15}$$

Conditional discovery efficiency. To distinguish feasibility from structural discovery, we measure the probability that a feasible generation discovers a previously unoccupied realized niche:

$$P(\text{new} \mid F=1) = \frac{\sum_{t=1}^{B} \mathbf{1}[F_p(x_t)=1,\ c_p(x_t) \notin \mathcal{O}_{t-1}]}{\sum_{t=1}^{B} F_p(x_t)}. \tag{16}$$

For this diagnostic, seeds with no feasible generations are omitted from the prompt-level mean; prompts with no defined seed are excluded from the paired analysis.

For each prompt, metrics are averaged over search seeds before computing paired method differences. Confidence intervals are obtained by paired bootstrap resampling over prompts with $10{,}000$ replicates (Koehn, 2004), and we report percentile-based 95% confidence intervals. Unless otherwise stated, we use search seeds 101, 202, and 303.

C Additional Quantitative Results

Representative task confidence. Using the soft task-confidence score in Eq. 13, we evaluate the selected archive representatives. Relative to BOP-Elites, AGSI shows a small decrease on spatial relations ($\Delta = -0.0169$, 95% CI $[-0.0326, -0.0050]$), while the difference on exact counting is not statistically resolved ($\Delta = -0.0240$, 95% CI $[-0.0503, 0.0043]$). Thus, the coverage gain is not explained by higher representative task-confidence.

Post-hoc representative preference audit. We evaluate representatives post hoc using HPSv2 (Wu et al., 2023) and ImageReward (Xu et al., 2023). For each method and realized niche, the representative is selected using the same frozen task-confidence rule; HPSv2 and ImageReward affect neither search nor representative selection. To control for differences in structural coverage, we compare AGSI and BOP-Elites only on realized niches discovered by both methods under the same prompt and search seed. Table 8 reports the resulting paired differences. None of the four comparisons is statistically resolved, providing no evidence of lower representative preference for AGSI within matched realized niches. HPSv2 and ImageReward are learned preference metrics rather than human judgments and are therefore used only as auxiliary diagnostics.

Table 9: **Matched diagnostics for AGSI and Proxy-Balanced.** Δ denotes AGSI minus Proxy-Balanced. For $P(\text{new} \mid F = 1)$, the comparison includes prompts for which the conditional probability is defined (97 Spatial and 77 Count prompts). All other statistics use the complete Full GenEval prompt sets.

Task	Metric	AGSI	Proxy-Bal.	Δ (95% CI)
Spatial	Cov_{32}	6.410	6.010	$+0.400\ [0.260, 0.540]$
	Feasibility	0.488	0.492	$-0.003\ [-0.011, 0.005]$
	$P(\text{new} \mid F = 1)$	0.346	0.310	$+0.036\ [0.023, 0.050]$
Count	Cov_{32}	6.258	5.454	$+0.804\ [0.588, 1.021]$
	Feasibility	0.299	0.340	$-0.041\ [-0.057, -0.025]$
	$P(\text{new} \mid F = 1)$	0.763	0.613	$+0.151\ [0.123, 0.179]$

Table 10: **Fixed-control latent optimization on Full GenEval spatial relations.** Noise Diffusion optimizes z_0 while retaining the frozen Sobol controls. The comparison is matched in returned-generation budget, not in inference compute.

Method	Cov_{16}	Cov_{32}	AUC/B	Feas.	UNet / image
Sobol + random z_0	4.15	5.21	3.76	0.512	30
Noise Diffusion	4.94	5.88	4.48	0.697	2580
AGSI	4.89	6.36	4.36	0.487	30†

†AGSI additionally evaluates $K = 32$ cheap control proxies per search step; only the selected control is passed to the generator.

D Mechanism Analyses and Ablations

D.1 Feasibility and New-Niche Discovery

Table 9 reports the full comparison between AGSI and Proxy-Balanced, separating realized coverage, feasibility, and conditional new-niche discovery.

AGSI achieves higher coverage with nearly unchanged feasibility on Spatial and lower feasibility on Count, while $P(\text{new} \mid F = 1)$ is higher on both tasks. These results indicate that the coverage gain is associated with more effective new-niche discovery rather than higher feasibility.

D.2 Inference-Time Latent Optimization

We test whether improving realization under fixed controls can substitute for structural search. On the same 100 GenEval spatial-relation prompts and search seed 101, Noise Diffusion (Miao et al., 2025) optimizes the initial noise z_0 while retaining the frozen Sobol control sequence. We compare it with the corresponding Sobol and AGSI runs over a generation budget of $B = 32$. Table 10 summarizes the resulting coverage, feasibility, and inference-cost statistics.

Noise Diffusion provides a controlled test of realization quality because it retains the same Sobol control sequence and changes only the diffusion noise optimization. Its substantial increase in feasibility ($0.512 \to 0.697$), together with higher coverage than standard Sobol, shows that improving the realization of fixed controls can itself aid structural discovery. However, this improvement does not translate directly into maximal repertoire coverage: despite substantially lower feasibility, AGSI reaches 6.36 niches at $B = 32$, exceeding Noise Diffusion by $+0.48$ niches (95% CI $[0.14, 0.80]$).

The two methods are not statistically resolved at $B = 16$ or in AUC/B, so the evidence does not indicate a uniform advantage for AGSI throughout the search trajectory. Rather, the final-budget result highlights a different role of structural search: while latent optimization improves how successfully a fixed control is realized, AGSI reallocates subsequent generations according to which realized niches remain undiscovered. These mechanisms are therefore complementary rather than interchangeable. Because Noise Diffusion requires 2580 UNet evaluations per returned image versus 30 for standard GLIGEN generation, this experiment isolates their behavior under a matched returned-generation budget and should not be interpreted as a compute-efficiency comparison.



Figure 5: **Additional realized structural repertoire for "a photo of a book above a laptop."** The comparison uses search seed 101 at $B = 64$. BOP-Elites, Proxy-Balanced, and AGSI discover $6/9$, $8/9$, and $9/9$ realized niches, respectively.

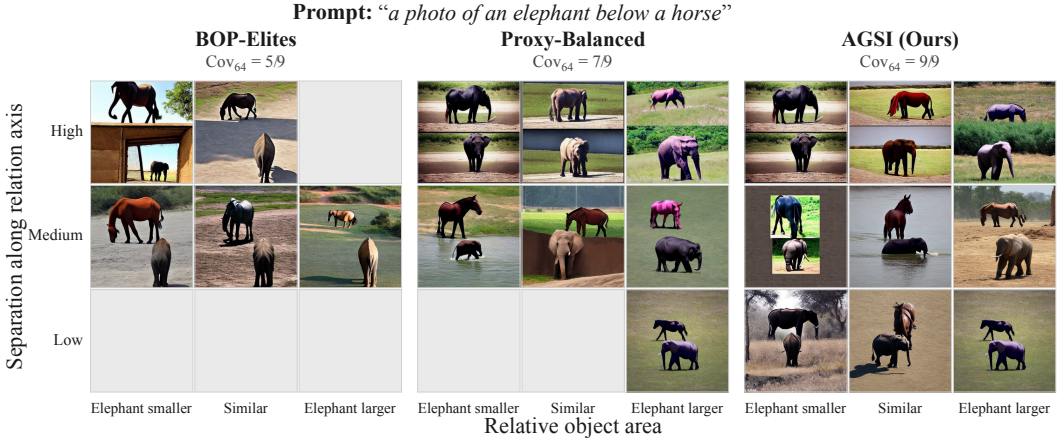

Figure 6: **Additional realized structural repertoire for "a photo of an elephant below a horse."** The comparison uses search seed 101 at $B = 64$. BOP-Elites, Proxy-Balanced, and AGSI discover $5/9$, $7/9$, and $9/9$ realized niches, respectively.

E Additional Qualitative Results

We provide additional qualitative examples for the primary Full GenEval spatial-relations setting, complementing the repertoire visualization in Fig. 3. Each figure compares BOP-Elites, Proxy-Balanced, and AGSI for the same prompt, search seed, and generation budget ($B = 64$). For each occupied realized niche j, we display the representative $x_j^\star(r) = \arg\max_{x \in \mathcal{H}_{j,B}} r(x)$, using the frozen soft task-confidence $s_p(x)$ in Eq. 13 as the representative utility for all methods. This post-hoc selection does not affect archive occupancy or the search trajectory. Rows encode separation along the requested relation axis, while columns encode smaller, similar, and larger relative area of the first object; gray cells denote undiscovered realized niches.

Figures 5 and 6 illustrate cases in which proxy balancing already provides a substantial improvement over BOP-Elites, while realized archive feedback further expands the repertoire. In both examples, BOP-Elites leaves several niches undiscovered, particularly at low relation-axis separation. Proxy-Balanced reaches more of these regions but still leaves one or two realized niches uncovered, whereas AGSI completes the 3×3 archive. These examples visually illustrate the distinction

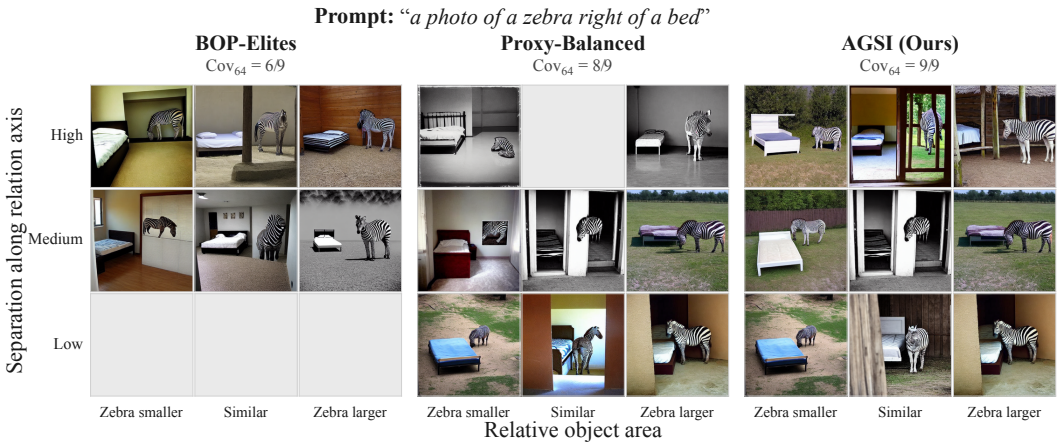

Figure 7: **Additional realized structural repertoire for "a photo of a zebra right of a bed."** The comparison uses search seed 303 at $B = 64$. BOP-Elites, Proxy-Balanced, and AGSI discover $6/9$, $8/9$, and $9/9$ realized niches, respectively.

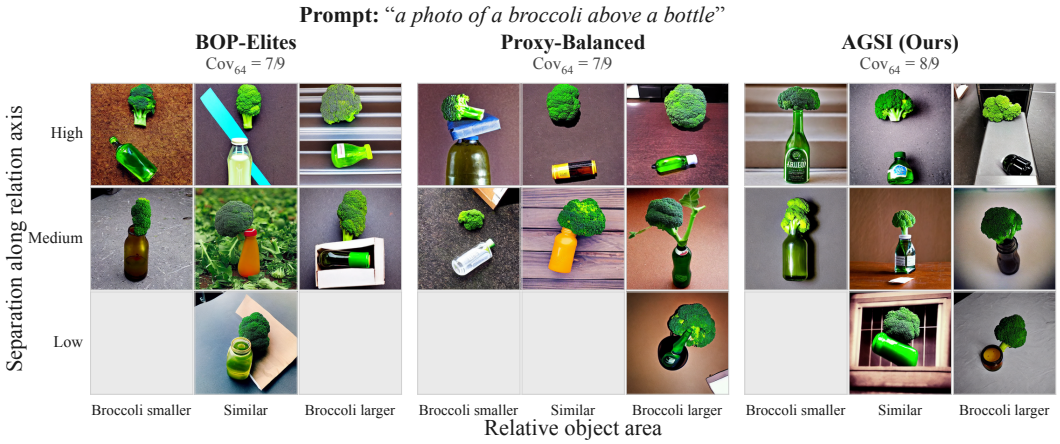

Figure 8: **Additional realized structural repertoire for "a photo of a broccoli above a bottle."** The comparison uses search seed 303 at $B = 64$. BOP-Elites and Proxy-Balanced each discover $7/9$ realized niches, while AGSI discovers $8/9$.

between balancing intended structural targets and allocating subsequent generations according to structures that have actually been realized.

The remaining examples show that the magnitude of the advantage varies across prompts. For the zebra–bed prompt in Fig. 7, AGSI again completes the archive, whereas Proxy-Balanced leaves one niche undiscovered. For the broccoli–bottle prompt in Fig. 8, the difference is smaller: AGSI discovers $8/9$ niches compared with $7/9$ for both BOP-Elites and Proxy-Balanced, and the low-separation, smaller-relative-area niche remains undiscovered. These examples therefore illustrate that realized archive guidance can further expand the repertoire, while not guaranteeing complete coverage for every prompt under a finite generation budget.

The qualitative results also expose a limitation of the current objective. Some representatives, particularly in Fig. 6 and Fig. 8, exhibit visible perceptual imperfections, such as unnatural appearance, object geometry, or composition, despite satisfying the feasibility criterion and occupying distinct structural niches. This is consistent with our design: AGSI optimizes the discovery of feasible realized structures, while the displayed representative in each niche is selected using the frozen task-

confidence in Eq. 13, rather than a perceptual-quality objective. Structural coverage should therefore not be interpreted as uniformly improved image quality. Instead, our results suggest a complementary role for structural search and quality-oriented optimization: AGSI determines which feasible structural regions to discover, while perceptual or realization-oriented methods could be applied to improve the representatives within those regions. We leave their joint optimization to future work.