Semantic Browsing: Controllable Diversity for Image Generation

Sara Dorfman*, Maya Vishnevsky*, Omer Dahary, Or Patashnik, and Daniel Cohen-Or

Tel Aviv University, Israel
*Equal contribution

Fig. 1: Semantic Browsing for Image Generation. From a single text prompt *A lion and a tiger in sunglasses, eating cotton candy*, the system produces a structured gallery of images that explore different meaningful interpretations of the same scene. Rather than random variations, each image reflects a distinct, coherent semantic choice (e.g., changes in character, interaction, or atmosphere) allowing users to browse a space of alternatives in a deliberate and interpretable way. In this visualization, the leftmost image serves as the root for the four variations in the center. The variation highlighted with an orange border is then selected as the specific parent for its four children displayed on the right.

Abstract. Modern text-to-image models excel in visual fidelity and prompt adherence. However, this strict adherence comes at the cost of diversity: generated samples tend to collapse into a single visual interpretation. Existing methods to improve diversity produce outputs driven by incidental variations rather than meaningful design choices. This motivates a new variant of the diversity task where structure is enforced on the generated samples.

We introduce a method for controlled diversity that enables *Semantic Browsing*, where users can navigate structured image galleries and experience creative exploration through a systematic traversal of meaningful, interpretable axes of variation. Achieving this level of semantic control requires a deep understanding of the scene. We exploit the fact that recent text-to-image models are trained on elaborated captions, effectively decoupling semantic decision-making from pixel generation. This enables a paradigm shift: instead of relying on stochastic variation within the text-to-image model, we induce diversity directly at the text level. By leveraging rich textual representations, we allow a Vision Language Model (VLM) to operate on the full scene context. To overcome the generic outputs typical of standard VLMs, we employ an *agentic workflow* that explicitly enforces structured variation attuned to the original prompt. We demonstrate that our method produces diverse and navigable design spaces where every variation corresponds to a specific, user-understandable semantic decision.

1 Introduction

Advancements in generative image models have rapidly transformed the way visual content is created, edited, and explored [9, 19, 35]. Much of the progress in these models has focused on visual fidelity and adherence to input conditioning. However, as these models become more capable, user expectations have expanded: rather than seeking a single faithful rendering, users often wish to explore multiple plausible outputs, particularly when their desired outcome is still unclear. This change in user expectations raises the challenge of generating a diverse gallery of outputs from a single input prompt.

Achieving such diversity is challenging, as recent state-of-the-art text-to-image models often exhibit limited semantic variation across samples generated from the same prompt (Figure 2). In particular, even when prompts are under-specified, different generations tend to converge on the same high-level semantic interpretation, differing only in visually insignificant details, or exhibiting severe biases [4]. A likely contributing factor to this lack of semantic diversity is the training paradigm of modern text-to-image models, which emphasizes strict adherence to highly detailed captions [1, 2, 17]. While this design choice substantially improves controllability and prompt faithfulness, it also biases the model toward committing to a single realization of the prompt, leaving little room for semantically diverse outputs.

Prior work has addressed this limitation by perturbing the conditioning signal [36, 41], introducing repulsive forces between sampling trajectories [6, 8], or generating large candidate pools from which diverse subsets are selected [31]. While successful at increasing diversity, these approaches do not offer explicit user control over the nature of the resulting variations. Consequently, differences across samples are driven by stochastic effects rather than explicit semantics.

In this work, we introduce the task of controlled semantic diversity, which enables users to explore generated images through meaningful, interpretable variations rather than relying on stochastic sampling. We refer to this process as *Semantic Browsing*, and view it as a conceptually different approach to diversity, where variations are explicitly specified rather than emergent. By semantic variations, we refer to changes in interpretable attributes of the image, such as object attributes or configurations (e.g., pose or spatial arrangement), global appearance factors (e.g., style, color palette, or lighting), or contextual elements (e.g., weather or background), while preserving all other aspects of the prompt, see Figure 1.

To achieve this controlled diversity, we impose structure on the diversity of generated outputs by leveraging the semantic reasoning capabilities of modern VLMs. Specifically, we introduce an agentic workflow that expands the user prompt into a richer semantic representation and identifies meaningful dimensions along which variation is both plausible and under-specified. These dimensions capture alternative semantic interpretations or design choices that are compatible with the original prompt but not explicitly specified by it. We then organize them into a structured set of prompt alternatives, each corresponding to a distinct semantic choice.

This prompt-based formulation places two key requirements on the underlying image generator. First, the generator must support fine-grained prompt-level control, so that semantic changes specified by the agentic workflow result in correspondingly precise visual changes. Second, it must preserve all aspects of the image that are not explicitly modified, ensuring that differences across the generated gallery arise solely from the intended semantic variations. Notably, recent state-of-the-art text-to-image models naturally satisfy these requirements, as they are trained for strict adherence to detailed textual specifications [2,17]. This makes them well suited to accurately reflect explicit prompt changes while maintaining consistency in attributes that are not mentioned. This training paradigm is exemplified by FIBO [17], which trains a text-to-image generator on long, structured captions to improve prompt adherence and controllability.

We evaluate our approach through extensive experiments across state-of-the-art text-to-image models, demonstrating consistent and substantial improvements in diversity over prior methods. Beyond increasing diversity, our method enables explicit control over the variations, allowing semantic differences to be specified and explored systematically rather than emerging from stochastic sampling. As illustrated in Figure 1, this results in structured galleries of images in which each output corresponds to a distinct, interpretable semantic alternative.

Fig. 2: Diversity Collapse in Standard Sampling. Prompt: *"A clown and a princess holding a wand."* While simply changing the random seed (bottom row) results in repetitive layouts [7] and limited variation, our method (top row) achieves significant structural and semantic diversity.

2 Related Work

Diversity in Text-to-Image Generation. Maintaining output diversity in Text-to-Image (T2I) systems is a persistent challenge, as common techniques like Classifier-Free Guidance (CFG) [20] often prioritize aesthetic fidelity at the cost of variety. Recent work [21] investigates the stage-wise dynamics of CFG, demonstrating how it suppresses diversity. This diversity collapse is further compounded in fast distilled diffusion models, a phenomenon directly linked to early generation dynamics [13].

To mitigate the CFG trade-off, Autoguidance [22] replaces the unconditional model in CFG with a weaker variant, effectively restoring diversity while maintaining image quality. However, this requires the computationally intensive training of a separate weak model. Although recent works [16,47] propose lightweight alternatives to address this burden, the approach has demonstrated limited reliability in practice.

CADS [36] and Guidance Interval [24] modulate the conditioning signal during denoising. While these methods improve sample variety, they can significantly degrade prompt alignment by relaxing guidance constraints. Other approaches, such as Particle Guidance [6] and MinorityPrompt [41], manipulate the sampling

process through latent repulsion or loss-based optimization at the latent level. However, because these methods operate primarily in the latent space, they lack the semantic granularity necessary for rich conceptual variety. Similarly, SGI [31] starts with a large pool of initial seeds and filters them during generation to reduce redundancy. While effective for batch variety, SGI is ultimately limited by the intrinsic diversity of the base generative model. Contextual Repulsion [8] addresses these limitations by applying repulsion in contextual attention space, improving semantic awareness and sample variety. However, it remains stochastic and lacks explicit control over specific semantic axes.

A more recent approach to prompt-level variety is PAG [48], which utilizes GFlowNets for diverse sampling. However, PAG is constrained by its reliance on a specific training dataset and lacks a global view over the relationships between generated prompts. In contrast, our approach is training-free and utilizes a hierarchical tree structure of generated images. By reasoning about multiple nodes collectively within the tree, our system ensures semantic diversity through structural inheritance, avoiding the repetitive results that often occur in independent or unstructured generation.

Beyond text-to-image generation, meaningful diversity [3] and hierarchical exploration [29] have also been studied in image restoration; our work instead uses hierarchy to organize explicit semantic alternatives for generation.

Creative Generation and Exploration. Our framework operates at the intersection of structured diversity and open-ended creative exploration, drawing on prior work in visual exploration and creativity-support tools [5, 23, 39, 46]. While methods like ConceptLab [33] and adaptive negative prompting [14] focus on exploring creative sub-categories of single objects, other approaches decompose and merge existing visual concepts for inspiration [10, 15, 34, 42]. In contrast, our method explicitly explores creative variations within the semantic space itself to organize alternative directions for generation.

Multi-Agent Systems for Controllable Generation. Current research has increasingly focused on utilizing specialized agents to enhance user control and refine the generation process. Maestro [43] employs a self-improving loop where multimodal agents act as critics to identify visual under-specification and iteratively refine the output for higher precision. Similarly, PromptSculptor [45] is a multi-agent framework that decomposes complex user queries into detailed, semantically rich descriptions to ensure the model captures every aspect of the user's intent. Proactive T2I Agents [18] further improve control by leveraging belief graphs to actively clarify ambiguous instructions through dialogue. TwinCo [44] follows a comparable strategy, employing an agentic feedback loop to systematically eliminate uncertainty in the prompt.

While these agent-based systems significantly enhance intent alignment and visual fidelity, they are fundamentally designed to converge on a single "best" version of the user's prompt. Since they focus on maximizing control over one optimal result, they ignore the many different ways a prompt could be interpreted.

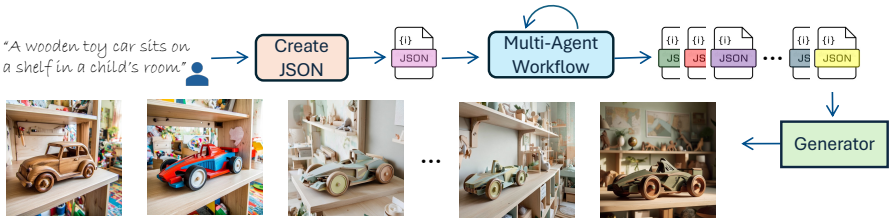

Fig. 3: Overview of the iterative generation flow. A user prompt is transformed into a structured JSON format which is iteratively modified by a Multi-Agent workflow. This process creates structured diversity of JSON variations that remain faithful to the initial user intent, driving the generator to produce perceptually distinct images.

Our work departs from these by using multiple agents to drive exploration instead of just narrowing down intent. By organizing generations into a hierarchical tree, we ensure the system produces a wide range of creative results rather than settling on a single interpretation.

3 Method

To enable controlled semantic exploration, we formalize the generation process as the construction of a hierarchical interpretative tree within a structured scene space. An overview of our method is demonstrated in Figure 3, with a concrete example of a generated tree shown in Figure 4. This section details our notation, the fundamental requirements for navigable diversity, and the multi-agent workflow that iteratively expands this tree through reasoned semantic refinements.

3.1 Setting

Our method operates within the space $\mathcal{S}$ of fully specified scene interpretations, encoded as structured JSONs. This format allows for fine-grained control over objects, attributes, and global scene properties. Given a user prompt p, we first expand it into an initial scene interpretation $s_0 \in \mathcal{S}$ using a VLM. This root scene represents one complete, plausible specification of the prompt. The output of our method is a rooted tree (V, E), where each node is a scene interpretation $s \in V \subset \mathcal{S}$.

In this structure, edges represent the atomic unit of semantic exploration. For any edge $(s_1, s_2) \in E$, there exists a semantic constraint c that transforms s_1 into s_2. Each constraint c is defined to be a specific instantiation of a broader semantic aspect a (e.g., *subject interactions*, *scene composition*, or *style*). For example, given a root scene s_0 derived from the prompt "A dog, a cat and a parrot" (Figure 4), a constraint c might instruct that the animals' *Interactions* are depicted as *Lively play*. This results in a child s_1 adhering to this behavior while preserving the remaining context of s_0. Practically, this transition is executed by a VLM-based scene refiner R such that $s_1 = R(s_0, c)$, ensuring that every step in the tree is both traceable and grounded in the preceding scene.

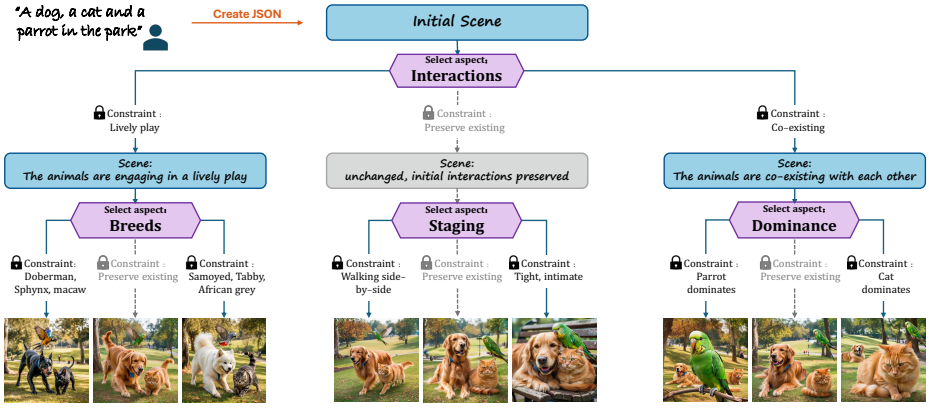

Fig. 4: Example of semantic browsing produced by our method. Starting from an initial scene interpretation inferred from the user prompt, the method explores alternative realizations by committing explicit semantic constraints at each step. Each branching point corresponds to alternative realizations of a single semantic aspect, while previously fixed constraints are preserved. Branching points also include an option to preserve the current value of the selected aspect, allowing exploration to continue along other semantic dimensions. Every node is a fully specified, renderable scene; 'preserve' branches propagate these states to the final level, ensuring the leaf nodes contain all generated representations ready for rendering.

Subsequently, we can render each node s using a modern prompt-adherent generator to produce a tree of images, enabling structured Semantic Browsing (Fig. 4).

3.2 Tree Requirements

To ensure the tree remains both diverse and navigable, we require that for any node s with a set of children, the applied set of semantic constraints must satisfy three interdependent requirements: (i) *Semantic Structuring*: All children of a parent node must be derived from a shared semantic aspect a. This property is essential for structured browsing, as it ensures that the branching at each level explores variations along a single, semantically meaningful dimension. For instance, in Figure 4, the children of the root node vary strictly based on the *Interactions* between the animals, while the children of the rightmost branch vary based on the *Dominance* in the scene. (ii) *Heterogeneity*: Each constraint c must realize the common aspect a in a unique manner. For example, under the *Interactions* aspect shown in Figure 4, one branch instantiates the scenario of *Lively play*, while its sibling instantiates a *Co-existing* dynamic. This is the primary driver of diversity, forcing the model to explore different conceptual directions within the same shared aspect. (iii) *Plausibility*: Each constraint c must be logically consistent with the original prompt p and the preceding constraints in its branch. Plausibility acts as a filter for Heterogeneity: it ensures that while branches differ, they remain faithful to the parent scene's established context.

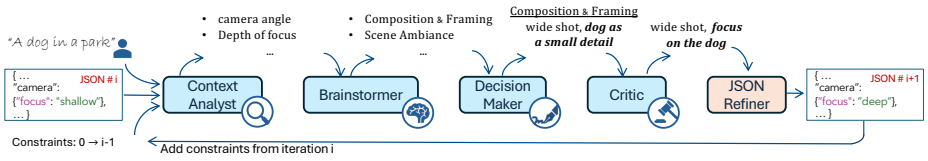

Fig. 5: Multi-Agent workflow guiding an iterative JSON generation process. The pipeline takes the current JSON configuration and a history of constraints derived from previous modifications (including the user prompt) as inputs. A sequence of agents—*Context Analyst*, *Brainstormer, Decision Maker,* and *Critic*—analyzes these inputs to select an aspect to modify and formulate specific instructions. The JSON Refiner then translates these instructions into an updated JSON configuration, and the new modifications are added to the constraint set for subsequent iterations.

Consider the rightmost branch in Figure 4: since it establishes that the animals are *Co-existing*, the subsequent *Cat dominates* constraint must be realized without aggression to avoid contradicting the parent state.

While these requirements define the target structure of the tree, balancing them simultaneously is a non-trivial reasoning task. We therefore employ a multi-agent workflow that serves as the engine for tree growth.

3.3 Agentic Workflow

Rather than generating the tree in a single pass, we expand it one node at a time. When the system expands a node s, our agentic workflow is triggered to generate its children through a staged process. The agentic workflow first identifies all details in the scene that remain flexible for change to ensure *Plausibility*, then combines these details into a single coherent aspect a to ensure *Structuring*, and finally proposes and critiques a set of candidate refinements to maximize *Heterogeneity*. This iterative, node-wise application ensures that every new set of children maintains the structural integrity and diversity required for effective Semantic Browsing.

Concretely, for every node s, we define the trajectory $C_s = (c_1, \ldots, c_n)$ as the ordered sequence of constraints applied along the path from the root s_0 to s. The workflow uses s, p, and C_s as context to ensure that new branches respect these previously fixed semantic decisions.

Next, we describe each component of the agentic workflow in detail. An illustration of their interactions is shown in Figure 5.

Context Analyst. The Context Analyst is tasked with defining the admissible search space for modification by identifying granular, low-level details, directly addressing the *Plausibility* requirement of the tree. It operates on the insight that a generated scene s is a composite of explicit specifications (enforced by the prompt p or the accumulated constraints C_s) and unconstrained details (filled in by the VLM to complete the scene), which we consider eligible for mutation. By explicitly distinguishing these, the Context Analyst isolates the set of mutable details $\{d_i\}$—such as specific colors, textures, or object subtypes—ensuring that subsequent changes target only the flexible components

of the scene without violating its established logical coherence. For example, in the scene from Figure 4, the Context Analyst identifies that while "a dog, a cat, and a parrot" must exist, their specific biological varieties (e.g., Doberman vs. Samoyed) are unconstrained details eligible for mutation. However, once a particular breed is added to the constraint set, the corresponding scene details become fixed for the rest of the subtree.

Brainstormer. The Brainstormer is responsible for laying the groundwork for meaningful *Semantic Structuring*, ensuring that the tree evolves through clear, meaningful concepts.

Given the initial prompt p and the accumulated constraints C_s, along with the set of low-level mutable details $\{d_i\}$ from the Context Analyst, the agent is tasked with identifying high-potential avenues for exploration. It applies inductive reasoning to synthesize semantic aspects $\{a_i\}$ by aggregating several low-level details into one high-level aspect. For instance, in the left branching in Figure 4, rather than varying the specific dog, cat, and bird species independently, the Brainstormer groups them under the cohesive aspect "Breeds," enabling coordinated modifications.

Crucially, it evaluates the potential of varying these candidates, explicitly assessing the magnitude of change (high, medium, or low) that varying each dimension would induce in the scene's *narrative*, *layout* and *style*. By prioritizing high-impact dimensions, the Brainstormer ensures that the tree evolves through significant conceptual shifts rather than trivial variations.

Decision Maker. The Decision Maker serves as the primary driver of *Heterogeneity*. By reasoning over the original prompt p, the current scene s, and the accumulated constraints C_s, the agent evaluates the candidate aspects $\{a_i\}$ suggested by the Brainstormer to identify prompt-dependent (see Supplementary Sec. 6) dimensions that offer the richest potential for variation. Operating strictly within this provided search space, it selects a single impactful dimension a^* and instantiates it into a set of alternative semantic constraints $\{c_i\}$. To ensure clear separation between sibling nodes, the Decision Maker actively reasons about the semantic boundaries of the scene, formulating constraints that offer widely divergent interpretations of a^* rather than incremental adjustments.

Critic. Finally, the Critic acts as the validation layer, primarily enforcing *Plausibility*. It reasons over the proposed constraints against the original prompt p and the accumulated constraints C_s, identifying potential contradictions or ambiguities that may have emerged during the creative process. The Critic validates that the proposals faithfully realize the intended concept while maintaining strict alignment with the prompt p and the accumulated context C_s. Aligning with self-correction strategies [11, 28], it then refines the candidate set into precise, executable instructions, ensuring that the final branches are not only semantically distinct but are robustly formulated to produce high-fidelity generations.

User Prompt: A group of people doing yoga.

User Prompt: A cat and a goldfish bowl.

Fig. 6: Structured diversity results. All images shown are derived from a single initial scene. Gray groups share a direct common ancestor, while colored boxes denote sibling branches: parallel variations with the same parent that differ by one semantic aspect. This demonstrates how our method introduces meaningful diversity while preserving the coherence of the original user prompt.

4 Experiments

In this section, we evaluate Semantic Browsing from three complementary perspectives. First, we demonstrate that our approach significantly enhances output diversity without compromising image quality or prompt alignment, benchmarking against established baselines designed to maximize diversity. Second, as Structured Diversity is a novel task whose hierarchical properties are not captured by existing diversity metrics, we introduce dedicated evaluations that measure the semantic and logical consistency of the generated hierarchy. Finally, we analyze the contribution of each component of our multi-agent workflow through ablation studies. Additional analyses, including a scaling ablation across tree depth and branching factor, as well as a sensitivity study of VLM choice, are provided in the Supplementary Material (Sections 7, 8).

Experimental Setup. We implement our method using the FIBO framework [17], leveraging its native prompt expander, refiner, and T2I generation modules. Additional details regarding the agent configuration and system prompts are provided in the Supplementary Material (Section 4).

Gallery Generation Strategy. To construct the final structured gallery, we employ our recursive tree expansion with a branching factor of three. At each node, the Decision Maker agent generates two distinct modification instructions, while a third branch retains the parent-node's JSON (identity mapping), ensuring that

Fig. 7: Qualitative comparison on the prompt: *"A glass bowl contains peeled tangerines and cut strawberries."* Columns 2 and 5-7 use consecutive seeds with hyperparameters optimized for diversity. Columns 3-4 display the most diverse subset of four images selected from a larger candidate pool. While baseline methods exhibit limited variation, our method (column 1) presents distinct and coherent interpretations. Examples include modifying the overall scene context, such as relocating the bowl to an outdoor picnic setting (row 1) or to a modern kitchen (row 4), the ordering and arrangement of the fruit (row 2), and the lighting conditions (row 3).

intermediate nodes are propagated unchanged to the leaf level. We expand this tree for three iterations, resulting in a final set of 27 leaf nodes (3^3), which constitutes our structured image gallery.

Baselines. To rigorously evaluate the effectiveness of our approach, we compare it against several methods. For fair comparison, all baselines were implemented using the same underlying generation model (FIBO), and their hyperparameters were optimized to maximize diversity specifically in this setting. Full implementation details are provided in the Supplementary Material (Section 3).

We employ three VLM-stochasticity-based baselines: *Stochastic VLM Seeding* generates the target gallery by simply varying the random seed of the VLM to leverage inherent model stochasticity; *Post-Hoc Diversity Optimization* applies a 'generate-and-select' strategy, filtering a pool of 79 candidates generated with different VLM seeds (strictly matching our method's total VLM call budget) to retain the subset that explicitly maximizes diversity; and *High-Temperature Post-Hoc Diversity Optimization*, which additionally increases sampling entropy of the VLM to force the selection of lower-probability tokens.

Table 1: Comparison to Baselines. Our method achieves top diversity (Vendi, DINO) while maintaining competitive Aesthetic scores; although lower on VQAScore, the result still reflects strong prompt adherence.

Method	Vendi ↑	DINO Sim. ↓	Aesthetic ↑	VQAScore ↑
Semantic Browsing (Ours)	**3.34**	**0.61**	6.52	0.90
Stochastic VLM Seeding	2.60	0.76	6.53	**0.93**
Post-Hoc Diversity Opt.	2.79	0.67	6.51	0.92
High-Temp. Post-Hoc Opt.	2.85	0.66	**6.56**	0.92
CADS	3.29	0.67	6.30	0.89
Guidance Interval	2.96	0.71	6.42	0.92
Power-Law CFG	2.75	0.74	6.28	0.93

Furthermore, we evaluate established generator-level methods that induce diversity directly within the denoising process: *CADS* [36], *Guidance Interval* [24], and *Power-Law CFG* [32]. To test whether these inference techniques provide additive diversity beyond simple random initialization, we applied them in conjunction with Stochastic VLM Seeding to generate the full gallery of 27 images.

4.1 Qualitative Evaluation

We begin by presenting a visual overview of our generated outputs in Figure 6, with additional examples shown in the Supplementary Material (Figures 6, 7). These examples demonstrate the framework's ability to synthesize a rich variety of semantic interpretations from a single input prompt, spanning the full spectrum from granular entity adjustments to holistic changes in setting and mood. Notably, the results are structured into triplets, where each group stems from a shared unique ancestor node, highlighting how early branching decisions propagate into distinct yet internally consistent variations. Crucially, this expansion in diversity does not come at the cost of visual quality; the generated images consistently exhibit high fidelity and aesthetic coherence, validating our approach's ability to balance broad semantic exploration with high-quality generation.

To validate diversity against existing baselines, Figure 7 compares results for the prompt: *"A glass bowl contains peeled tangerines and cut strawberries."* While baseline methods converge on a single mode, our approach uncovers distinct and plausible interpretations. As shown in the first column, our framework successfully modifies the overall scene context, such as relocating the bowl to an outdoor picnic setting (row 1) or to a modern kitchen (row 4). We also vary the ordering and arrangement of the fruit (row 2) and the lighting conditions (row 3). This confirms our ability to retrieve heterogeneous, high-fidelity alternatives without compromising prompt adherence. Additional qualitative comparisons are provided in the Supplementary Material (Figures 3-5).

4.2 Quantitative Evaluation

Dataset. We conduct our evaluation on a subset of 50 prompts randomly sampled from the MS-COCO dataset [26].

Metrics. To provide a comprehensive assessment of our method, we report metrics across three dimensions: diversity, image quality, and prompt adherence. *Diversity* is quantified via the Vendi Score [12] in Inception space [40] and pairwise DINO [30] similarity, capturing the extent of semantic variation across the gallery. To evaluate *quality* and validate that diversity-enhancing mechanisms do not degrade visual fidelity, we report the Aesthetic Score [37] (utilizing the LAION-based predictor [38]). For *prompt adherence* we utilize VQAScore [27].

Table 1 presents the quantitative results against all baselines. Our method achieves superior diversity, securing the highest Vendi Score (3.34) and the lowest DINO Similarity (0.61), significantly outperforming all competing approaches. Crucially, this substantial expansion in semantic coverage is achieved while maintaining comparable Aesthetic Scores (6.52). This confirms that our framework successfully balances high-variance exploration with high image quality, avoiding the significant degradation often associated with maximizing diversity. While we observe a decrease in VQAScore, this may reflect inherent model biases within the evaluators, which often favor conventional, low-variance compositions rather than a true decline in prompt adherence. Regardless, the observed difference remains practically negligible.

We additionally report the computational overhead of our agentic workflow in the Supplementary Material (Section 5). Despite the added structure, Semantic Browsing remains competitive in cost with baseline methods.

User Study. Standard quantitative metrics often rely on dataset biases that penalize the very diversity our method aims to achieve. To directly assess perceptual quality and diversity, we conducted a head-to-head human evaluation with 25 participants, utilizing 12 randomly selected prompts for each baseline comparison. We compared Semantic Browsing against CADS, Guidance Interval, Power-Law CFG, and Post-Hoc Optimization. For the Post-Hoc baseline, we used the High-Temperature configuration as it yielded the optimal metric performance in Table 1. As shown in Figure 8, our method outperforms all baselines, achieving substantial win rates for superior diversity while securing the majority vote for overall preference.

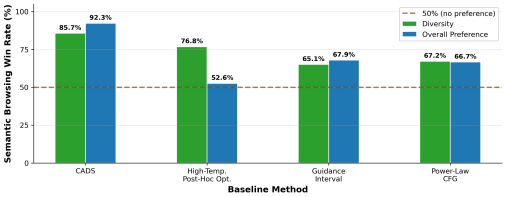

Fig. 8: Human Preference Study. Our method (Semantic Browsing) dominates in Diversity across all comparisons while consistently outperforming baselines in Overall Preference.

4.3 Structure Evaluation

Since Structured Diversity is a novel task, standard metrics are ill-equipped to capture the relational properties of the generated gallery. While adequate for quantifying the Heterogeneity requirement (Section 3.2), these metrics treat outputs as independent samples, ignoring the hierarchical dependencies that are unique to our method. Therefore, we introduce two evaluation protocols to specifically validate structural integrity and logical consistency.

For these structural evaluations, we deviate from the gallery generation process described previously. Instead of inspecting only final leaf nodes (which include identity-mapped copies), we evaluate the complete set of unique nodes in the tree to accurately assess the internal coherence of the generation hierarchy.

Semantic-Topological Correlation. We hypothesize that the semantic distance between two images should correlate with their topological distance in the generation tree. This relationship is a direct consequence of the *Semantic Structuring* requirement (Section 3.2), which enforces that exactly one aspect changes between parent and child nodes. To verify this, we analyze Pairwise DINO Distance as a function of graph distance (path length between nodes). Figure 9 presents the results of this analysis, aggregated across 50 generated trees (17,550 total pairs). We observe a strong positive correlation between the topological distance in the tree and the semantic distance in the image space. As the number of graph hops between two nodes increases, the median pairwise DINO distance rises monotonically (ranging from 0.168 at 1 hop to 0.452 at 5 hops). This confirms that our framework successfully satisfies the *Semantic Structuring* requirement; the hierarchy effectively encodes semantic relationships, where neighboring nodes share visual characteristics and distant nodes exhibit greater semantic divergence.

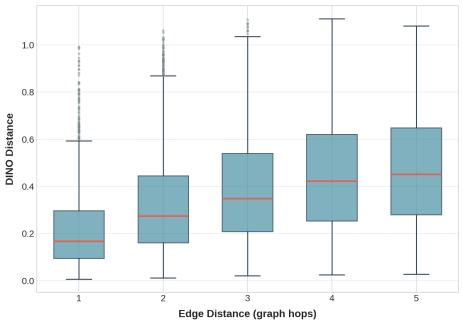

Fig. 9: Semantic-Topological Correlation. Box plot showing the distribution of Pairwise DINO Distances as a function of graph distance (number of edge hops between nodes). The clear upward trend validates that our generation tree creates a coherent semantic space, where topological proximity translates to semantic similarity.

Hierarchical Consistency. To validate that the tree maintains logical continuity, we utilize LLM-as-a-judge [25] to measure the alignment between a generated node and the constraints inherited from its ancestors. This metric explicitly validates the *Plausibility* requirement (Section 3.2) by ensuring that diversity modifications do not violate established context. Our framework achieves a high consistency score of 0.87 (out of 1.0), demonstrating that in the vast majority of cases, generated nodes successfully adhere to the cumulative constraints derived from the full root-to-node path. We note that this metric penalizes only the first violation of a constraint; we do not cumulatively penalize a child node if it remains consistent with a parent that has already violated a constraint, allowing us to isolate exactly where divergences occur.

4.4 Ablation Study

To validate the architectural design of our framework, we conduct an ablation study to isolate the specific contribution of each agent. Our analysis confirms that

the multi-agent decomposition is essential, as each component plays a distinct and necessary role in the generation pipeline. To verify that constraints are not violated—an essential aspect of Plausibility—we utilize a VLM-as-a-judge [25] to measure the alignment between a generated node and the constraints inherited from its ancestors. We refer to the average of this score as *Hierarchical Consistency*.

Context Analyst. When the Context Analyst is removed, the burden of interpreting the semantic gap between the high-level user prompt and the low-level JSON scene representation falls entirely on the Brainstormer. Without explicitly enforcing the internalization of this gap, non-admissible details change, resulting in a significant drop in Plausibility. Table 3 (w/o Context Analyst) quantifies this impact. While the VQAScore remains stable (0.90), indicating that prompt adherence is preserved, the Hierarchical Consistency drops from 0.87 to 0.82. This divergence confirms that the Context Analyst is specifically required to maintain contextual continuity, directly associated with the Plausibility requirement.

Brainstormer and Decision Maker. We evaluate the impact of merging the Brainstormer and Decision Maker into a single, unified agent. Since the Brainstormer is responsible for Semantic Structuring and the Decision Maker for Heterogeneity, the unified agent struggles to optimize for both tasks simultaneously, leading to a degradation in tree quality. Separating these roles increases the global mean DINO distance from 0.362 (unified) to 0.389 (separated), representing a 7.2% relative improvement in overall diversity.

Table 2: Brainstormer / Decision Maker Ablation. Comparison of mean pairwise DINO distance. The full workflow consistently yields higher diversity across all graph hops.

Edge Dist.	Full	Unified
1	0.221	0.218
2	0.326	0.313
3	0.392	0.365
4	0.450	0.411
5	0.473	0.439

Table 2 decomposes this improvement by topological distance. The full workflow consistently exhibits larger DINO distances across all edge distances, confirming that the specialized role separation yields significantly greater structured diversity compared to the unified ablation.

This demonstrates the critical roles of these agents in maintaining Heterogeneity and Semantic Structure, validating that distinct architectural components are required to satisfy these dual objectives.

Critic. The Critic acts as the final safeguard for prompt adherence and logical consistency. Table 3 (w/o Critic) shows that ablating this agent reduces VQAScore from 0.90 to 0.87, while Hierarchical Consistency remains stable. This suggests that while upstream agents mostly maintain internal constraint consistency, the Critic remains a necessary component to catch rare violations that do oc-

Table 3: Ablation of Agents Responsible for Plausibility. The Context Analyst improves logical continuity (Hierarchical Consistency), while the Critic preserves prompt adherence (VQAScore). Together, they maintain the full scope of plausibility.

Metric	Full Workflow	w/o Context Analyst	w/o Critic
VQAScore	0.90	0.90	0.87
Hierarchical Consistency	0.87	0.82	0.87

cur. The drop in VQAScore confirms that the Critic is essential for preventing semantic drift, ensuring that the output remains prompt adherent.

5 Conclusions, Limitations and Future Work

We have presented a structured approach for generating semantic diversity in text-to-image models. At a high level, this work adopts a perspective in which diversity arises from explicit semantic decision-making rather than from stochastic variation alone. By committing to concrete semantic choices during generation, differences between outputs become interpretable and persistent rather than incidental. Consequently, the generated results form not just a collection of images, but a structured and navigable semantic space of alternative scene interpretations.

This perspective was enabled by recent text-to-image models that exhibit strong prompt adherence, which we treated not as a limitation on diversity but as an enabler for precise semantic control. Rather than optimizing toward a single refined output, the formulation emphasized exploration, using a multi-agent reasoning process to surface and maintain multiple plausible interpretations of an under-specified prompt. These interpretations were constructed through sequences of inherited semantic commitments, leading to structured semantic variation in which previously fixed decisions remained consistent while new variations were introduced in a controlled and interpretable manner.

The method presented here has several limitations that stem primarily from its current realization. The quality and usefulness of the explored semantic space depend on the underlying generative model's ability to faithfully realize fine-grained prompt modifications, as well as on the semantic reasoning capabilities of the agents that propose and evaluate variations. In particular, while modern VLMs are effective at maintaining consistency and plausibility, their ability to propose rich and diverse semantic alternatives remains limited relative to the breadth of interpretations one might ultimately wish to explore, which can constrain the scope of the resulting semantic space.

More broadly, although this work focused on image generation, the notion of structuring diversity through explicit semantic decisions suggests a more general paradigm. Looking forward, we believe that structured semantic exploration could extend beyond images to other generative domains, such as video, 3D content, or multimodal generation, offering a principled way to move from isolated outputs toward coherent, navigable spaces of alternatives.

Acknowledgments

We thank Nir Goren, Saar Huberman, and Shelly Golan for helpful discussions and early feedback on this work. We also thank the ECCV 2026 reviewers for their constructive comments and suggestions. This research was supported in part by the Israel Science Foundation (grants no. 2492/20 and 1473/24), Len Blavatnik, and the Blavatnik Family Foundation. We also thank NVIDIA for their generous support through the NVIDIA Academic Grant Program, which provided GPU hours via Brev for this research.

References

1. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. vol. 2, p. 8 (2023)
2. Black Forest Labs: Flux, https://github.com/black-forest-labs/flux (2024), https://github.com/black-forest-labs/flux
3. Cohen, N., Manor, H., Bahat, Y., Michaeli, T.: From posterior sampling to meaningful diversity in image restoration. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 6407–6444 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/19e2ed0e9f1a21bef660c7f83742ef56-Paper-Conference.pdf
4. Cohen, N., Spingarn-Eliezer, N., Huberman-Spiegelglas, I., Michaeli, T.: Minethegap: Automatic mining of biases in text-to-image models (2025), https://arxiv.org/abs/2512.13427
5. Cohen-Or, D., Zhang, H.: From inspired modeling to creative modeling. Vis. Comput. **32**(1), 7–14 (Jan 2016). https://doi.org/10.1007/s00371-015-1193-9, https://doi.org/10.1007/s00371-015-1193-9
6. Corso, G., Xu, Y., De Bortoli, V., Barzilay, R., Jaakkola, T.: Particle guidance: non-iid diverse sampling with diffusion models. arXiv preprint arXiv:2310.13102 (2023)
7. Dahary, O., Cohen, Y., Patashnik, O., Aberman, K., Cohen-Or, D.: Be decisive: Noise-induced layouts for multi-subject generation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025)
8. Dahary, O., Koren, B., Garibi, D., Cohen-Or, D.: On-the-fly repulsion in the contextual space for rich diversity in diffusion transformers (2026), https://arxiv.org/abs/2603.28762
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021)
10. Dorfman, S., Cohen-Bar, D., Gal, R., Cohen-Or, D.: Ip-composer: Semantic composition of visual concepts. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. SIGGRAPH Conference Papers '25, Association for Computing Machinery, New York, NY, USA (2025), https://doi.org/10.1145/3721238.3730624
11. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate (2023), https://arxiv.org/abs/2305.14325
12. Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning (2023), https://arxiv.org/abs/2210.02410
13. Gandikota, R., Bau, D.: Distilling diversity and control in diffusion models (2025), https://arxiv.org/abs/2503.10637
14. Golan, S., Nitzan, Y., Wu, Z., Patashnik, O.: Vlm-guided adaptive negative prompting for creative generation. arXiv preprint arXiv:2510.10715 (2025)
15. Goldberg, K., Richardson, E., Vinker, Y.: Inspiration seeds: Learning non-literal visual combinations for generative exploration. arXiv preprint arXiv:2602.08615 (2026)
16. Gu, E., Hou, H.: In-situ autoguidance: Eliciting self-correction in diffusion models (2025), https://arxiv.org/abs/2510.17136

17. Gutflaish, E., Kachlon, E., Zisman, H., Hacham, T., Sarid, N., Visheratin, A., Huberman, S., Davidi, G., Bukchin, G., Goldberg, K., et al.: Generating an image from 1,000 words: Enhancing text-to-image with structured captions. arXiv preprint arXiv:2511.06876 (2025)
18. Hahn, M., Zeng, W., Kannen, N., Galt, R., Badola, K., Kim, B., Wang, Z.: Proactive agents for multi-turn text-to-image generation under uncertainty. arXiv preprint arXiv:2412.06771 (2024)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
20. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
21. Jin, C., Shi, Q., Gu, Y.: Stage-wise dynamics of classifier-free guidance in diffusion models. arXiv preprint arXiv:2509.22007 (2025)
22. Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a diffusion model with a bad version of itself. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=bg6fVPVs3s
23. Kleiman, Y., Lanir, J., Danon, D., Felberbaum, Y., Cohen-Or, D.: Dynamicmaps: Similarity-based browsing through a massive set of images. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. p. 995–1004. CHI '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2702123.2702224, https://doi.org/10.1145/2702123.2702224
24. Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., Lehtinen, J.: Applying guidance in a limited interval improves sample and distribution quality in diffusion models. Advances in Neural Information Processing Systems **37**, 122458–122483 (2024)
25. Lee, S., Kim, S., Park, S.H., Kim, G., Seo, M.: Prometheus-vision: Vision-language model as a judge for fine-grained evaluation (2024), https://arxiv.org/abs/2401.06591
26. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015), https://arxiv.org/abs/1405.0312
27. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation (2024), https://arxiv.org/abs/2404.01291
28. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al.: Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems **36**, 46534–46594 (2023)
29. Nehme, E., Mulayoff, R., Michaeli, T.: Hierarchical uncertainty exploration via feedforward posterior trees. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 125142–125191. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-3975, https://proceedings.neurips.cc/paper_files/paper/2024/file/e262fc23ec7275230ee77c55d0cc9555-Paper-Conference.pdf
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2:

Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
31. Parmar, G., Patashnik, O., Ostashev, D., Wang, K.C., Aberman, K., Narasimhan, S., Zhu, J.Y.: Scaling group inference for diverse and high-quality generation. arXiv preprint arXiv:2508.15773 (2025)
32. Pavasovic, K.L., Verbeek, J., Biroli, G., Mezard, M.: Classifier-free guidance: From high-dimensional analysis to generalized guidance forms (2025), https://arxiv.org/abs/2502.07849
33. Richardson, E., Goldberg, K., Alaluf, Y., Cohen-Or, D.: Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. ACM Transactions on Graphics **43**(3), 1–14 (2024)
34. Richardson, E., Goldberg, K., Alaluf, Y., Cohen-Or, D.: Piece it together: Part-based concepting with ip-priors (2025), https://arxiv.org/abs/2503.10365
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022)
36. Sadat, S., Buhmann, J., Bradley, D., Hilliges, O., Weber, R.M.: Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. arXiv preprint arXiv:2310.17347 (2023)
37. Schuhmann, C.: Improved aesthetic predictor (2022), https://github.com/christophschuhmann/improved-aesthetic-predictor
38. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022), https://arxiv.org/abs/2210.08402
39. Shapira, L., Shamir, A., Cohen-Or, D.: Image appearance exploration by model-based navigation. Computer Graphics Forum **28**(2), 629–638 (2009). https://doi.org/https://doi.org/10.1111/j.1467-8659.2009.01403.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2009.01403.x
40. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision (2015), https://arxiv.org/abs/1512.00567
41. Um, S., Ye, J.C.: Minority-focused text-to-image generation via prompt optimization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20926–20936 (2025)
42. Vinker, Y., Voynov, A., Cohen-Or, D., Shamir, A.: Concept decomposition for visual exploration and inspiration (2023), https://arxiv.org/abs/2305.18203
43. Wan, X., Zhou, H., Sun, R., Nakhost, H., Jiang, K., Sinha, R., Arık, S.Ö.: Maestro: Self-improving text-to-image generation via agent orchestration. arXiv preprint arXiv:2509.10704 (2025)
44. Wang, J., He, Y., Zhong, Y., Song, X., Su, J., Feng, Y., Wang, R., He, H., Zhu, W., Yuan, X., et al.: Twin co-adaptive dialogue for progressive image generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3645–3653 (2025)
45. Xiang, D., Xu, W., Chu, K., Ding, T., Shen, Z., Zeng, Y., Su, J., Zhang, W.: Promptsculptor: Multi-agent based text-to-image prompt optimization. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 774–786 (2025)
46. Xu, K., Zhang, H., Cohen-Or, D., Chen, B.: Fit and diverse: set evolution for inspiring 3d shape galleries. ACM Trans. Graph. **31**(4) (Jul 2012). https://doi.org/10.1145/2185520.2185553, https://doi.org/10.1145/2185520.2185553

47. Yehezkel, S., Dahary, O., Voynov, A., Cohen-Or, D.: Navigating with annealing guidance scale in diffusion space. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
48. Yun, T., Zhang, D., Park, J., Pan, L.: Learning to sample effective and diverse prompts for text-to-image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23625–23635 (2025)