Semantic Browsing: Controllable Diversity for Image Generation

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

Tel Aviv University

Fig. 1. **Semantic Browsing for Image Generation.** From a single text prompt *"A poster featuring animals"*, 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, composition, or style) 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 a purple border is then selected as the specific parent for its four children displayed on the right.

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. *Project page: https://saradorfman1.github.io/SemanticBrowsing-webpage/*

1 Introduction

Advancements in generative image models have rapidly transformed the way visual content is created, edited, and explored [Dhariwal and Nichol 2021; Ho et al. 2020; Rombach et al. 2022]. 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 underspecified, different generations tend to converge on the same high-level semantic interpretation, differing only in visually insignificant details, or exhibiting severe biases [Cohen et al. 2025]. 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 [Betker et al. 2023; Black Forest Labs 2024; Gutflaish et al. 2025]. 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 [Sadat et al. 2023; Um and Ye 2025], introducing repulsive forces between sampling trajectories [Corso et al. 2023; Dahary et al. 2026], or generating large candidate pools from which diverse subsets are selected [Parmar et al. 2025]. 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 semantic factors.

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,

*Indicates equal contribution.

Fig. 2. **Diversity Collapse in Standard Sampling.** Visual comparison for the prompt: *"A clown and a princess holding a wand."* While simply changing the random seed (consecutive seeds 0-3 shown in bottom row) results in repetitive layouts [Dahary et al. 2025] and limited variation, our method (top row) achieves significant structural and semantic diversity.

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 [Black Forest Labs 2024; Gutflaish et al. 2025]. 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 [Gutflaish et al. 2025], 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 nature of 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.

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) [Ho and Salimans 2022] often prioritize aesthetic fidelity at the cost of variety. Recent work [Jin et al. 2025] 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 [Gandikota and Bau 2025].

To mitigate the CFG trade-off, Autoguidance [Karras et al. 2024] 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 [Gu and Hou 2025; Yehezkel et al. 2025] propose lightweight alternatives to address this burden, the approach has demonstrated limited reliability in practice.

CADS [Sadat et al. 2023] and Guidance Interval [Kynkäänniemi et al. 2024] 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 [Corso et al. 2023] and MinorityPrompt [Um and Ye 2025], 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 [Parmar et al. 2025] 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. To overcome these limitations, Contextual Repulsion [Dahary et al. 2026] applies repulsion within the contextual attention space. Although this shift improves semantic awareness and sample variety, the method still relies on stochastic diversity without explicit control over specific semantic axes.

A more recent approach to prompt-level variety is PAG [Yun et al. 2025], 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 [Cohen et al. 2024] and hierarchical exploration [Nehme et al. 2024] 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. While methods like ConceptLab [Richardson et al. 2024] and adaptive negative prompting [Golan et al. 2025] focus on exploring creative sub-categories of single objects, other approaches decompose and merge existing visual concepts for inspiration [Goldberg et al. 2026; Vinker et al. 2023]. In contrast, our method explicitly

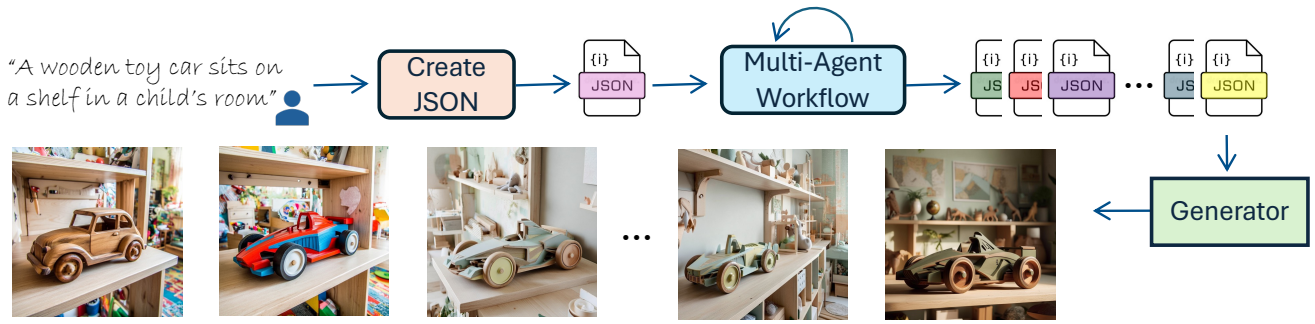

Fig. 3. **Overview of the iterative generation flow.** A user prompt is transformed into a structured JSON format which is iteratively modified by a Multi-Agent workflow. This process creates structured diversity of JSON variations that remain faithful to the initial user intent, driving the generator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

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 [Wan et al. 2025] 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 [Xiang et al. 2025] 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 [Hahn et al. 2024] further improve control by leveraging belief graphs to actively clarify ambiguous instructions through dialogue. Twin-Co [Wang et al. 2025] 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. 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.

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

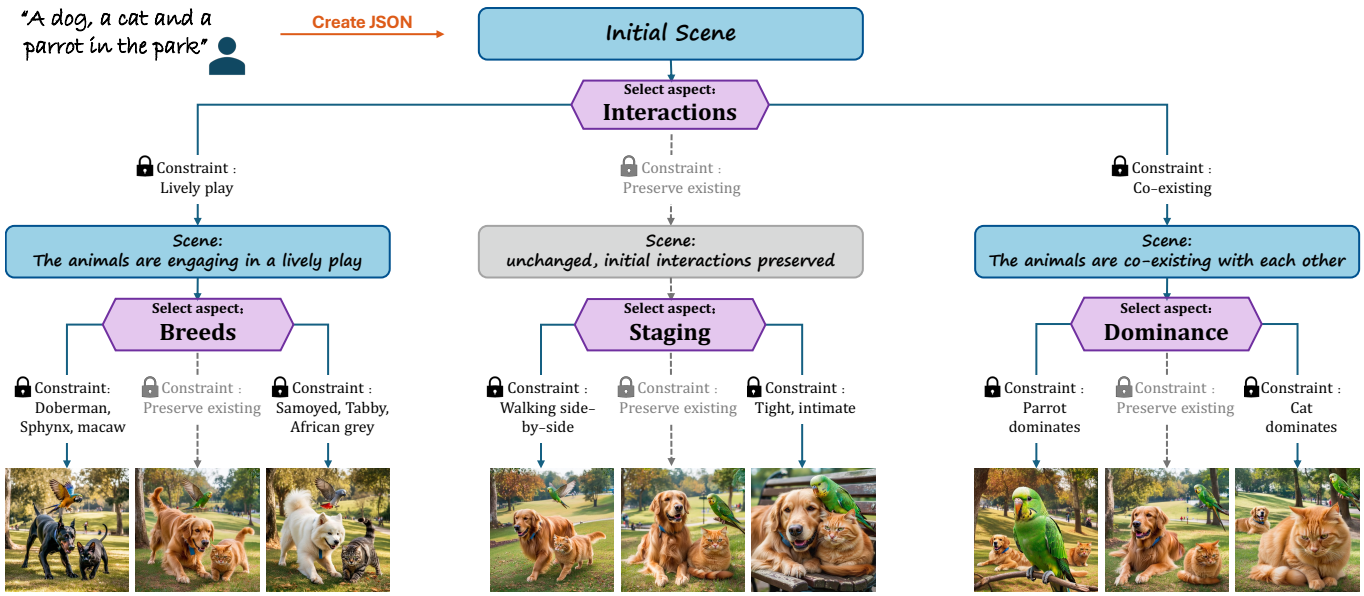

Fig. 4. **Example of semantic browsing produced by our method.** Starting from an initial scene interpretation inferred from the user prompt, the method explores alternative realizations by committing explicit semantic constraints at each step. Each branching point corresponds to alternative realizations 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.

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. 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 sub-types—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 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

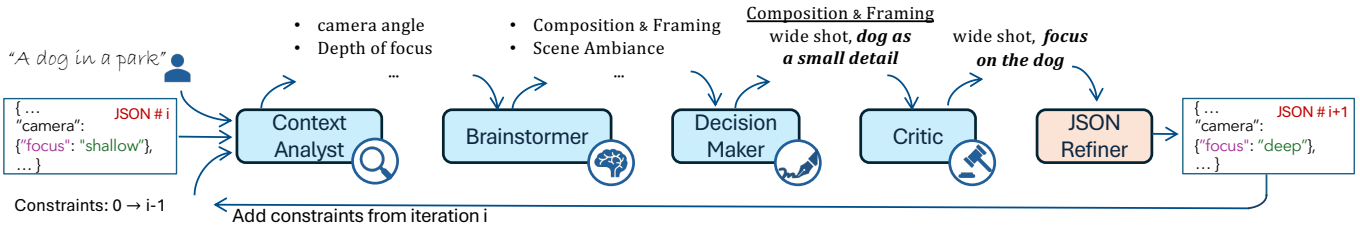

Fig. 5. **Multi-Agent workflow guiding an iterative JSON generation process.** The pipeline takes the current JSON configuration and a history of constraints derived from previous modifications (including the user prompt) as inputs. A sequence of agents—*Context Analyst*, *Brainstormer*, *Decision Maker*, 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.

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 Appendix D) 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 [Du et al. 2023; Madaan et al. 2023], 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.

Recent work demonstrates that prompting models to explicitly articulate their reasoning significantly enhances performance across various tasks [Wei et al. 2023; Yao et al. 2023]. Building on this literature, we design our agents to explicitly reason over their decisions before finalizing any action.

3.4 Interactive Browsing

Our design inherently supports *Interactive Browsing*: while we describe an automatic expansion strategy, the workflow allows a user

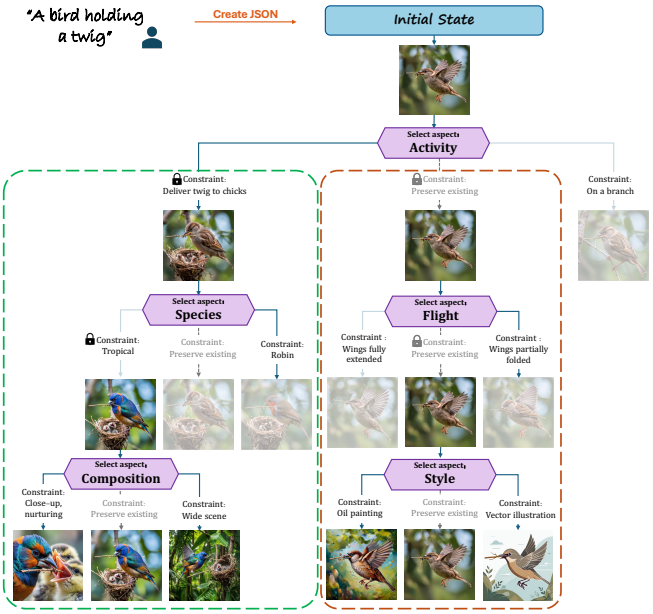

Fig. 6. **Example of interactive semantic browsing.** At each node, users may either commit to a new realization of the selected semantic aspect and continue refining that interpretation (green), or preserve the current realization and explore other semantic aspects from the same state (orange). All nodes correspond to valid intermediate states that can be further expanded.

to manually select any node of interest to trigger further generation, effectively continuing the exploration along a desired path, as demonstrated in Figure 6.

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 (Appendix F) and a sensitivity study of VLM choice (Appendix E), are provided in the appendix.

Experimental Setup. We implement our method using the FIBO framework [Gutflaish et al. 2025], leveraging its native prompt expander, refiner, and T2I generation modules. Additional details regarding the agent configuration and system prompts are provided in Appendix B.

Model-Agnostic Design. While our experimental results are obtained using FIBO's generation pipeline, the proposed framework itself is model-agnostic and decoupled from the underlying rendering backbone. To demonstrate this, we utilize FIBO's VLM-based modules for prompt enhancement and scene refinement, while employing a distinct architecture, FLUX.2 [Labs 2025], to render the final images. As shown in Figure 9, our framework successfully separates semantic control from rendering, ensuring consistent performance across different backbones.

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 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 Appendix A.

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.

Furthermore, we evaluate established generator-level methods that induce diversity directly within the denoising process: *CADS* [Sadat et al. 2023], *Guidance Interval* [Kynkäänniemi et al. 2024], and *Power-Law CFG* [Pavasovic et al. 2025]. 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 7, with additional examples shown in Figures 12 and 13 (Appendix). 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 8 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 Figures 14–16.

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

4.2 Quantitative Evaluation

Dataset. We conduct our evaluation on a subset of 50 prompts randomly sampled from the MS-COCO dataset [Lin et al. 2015].

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 [Friedman and Dieng 2023] in Inception space [Szegedy et al. 2015] and pairwise DINO [Oquab et al. 2024] similarity, capturing the extent of semantic variation across the gallery. To evaluate *quality* and validate diversity-enhancing mechanisms do not degrade visual fidelity, we report the Aesthetic Score [Schuhmann 2022] (utilizing the LAION-based predictor [Schuhmann et al. 2022]). For *prompt adherence* we utilize VQAScore [Lin et al. 2024].

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

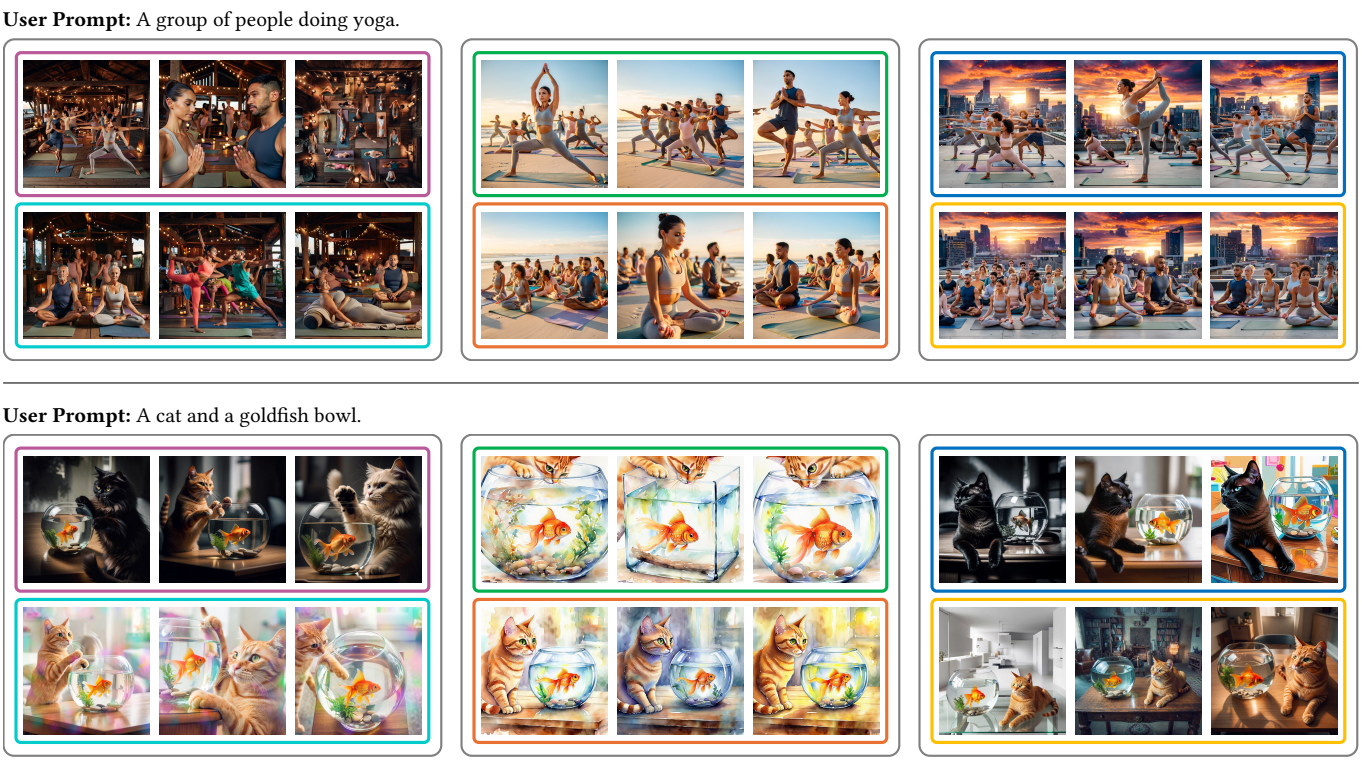

Fig. 7. **Structured diversity results.** All images shown are derived from a single initial scene. The outer gray groupings organize results that share a direct common ancestor scene. Inside, the colored boxes distinguish sibling branches (parallel variations that share the same parent but differ from one another by a single semantic aspect). This demonstrates how our method introduces meaningful diversity while preserving the coherence of the original user prompt.

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 Appendix C. 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 10, our method outperforms all baselines, achieving substantial win rates for superior diversity while consistently securing the majority vote for 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 the final leaf nodes (which include identity-mapped copies), we evaluate the complete set of unique nodes within 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 11 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*

Fig. 8. **Qualitative comparison on the prompt:** *"A glass bowl contains peeled tangerines and cut strawberries."* Columns 2 and 5-7 report results using 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) successfully 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).

Structuring requirement; the hierarchy effectively encodes semantic relationships, where neighboring nodes share visual characteristics and distant nodes exhibit greater semantic divergence.

Hierarchical Consistency. To validate that the tree maintains logical continuity, we utilize LLM-as-a-judge [Lee et al. 2024] 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 [Lee et al. 2024] 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

User Prompt: A dancer performing a dance.

User Prompt: A red fox and a white fox playing a video game.

Fig. 9. **Model-Agnostic Generation (FLUX.2).** Qualitative results demonstrating the transferability of our framework to the FLUX.2 architecture. By utilizing our agentic flow solely for scene generation and FLUX.2 as the rendering backbone, we achieve consistent structured diversity.

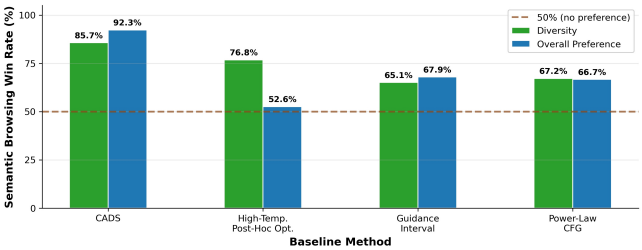

Fig. 10. **Human Preference Study.** Our method (Semantic Browsing) dominates in Diversity across all comparisons while consistently outperforming baselines in Overall Preference.

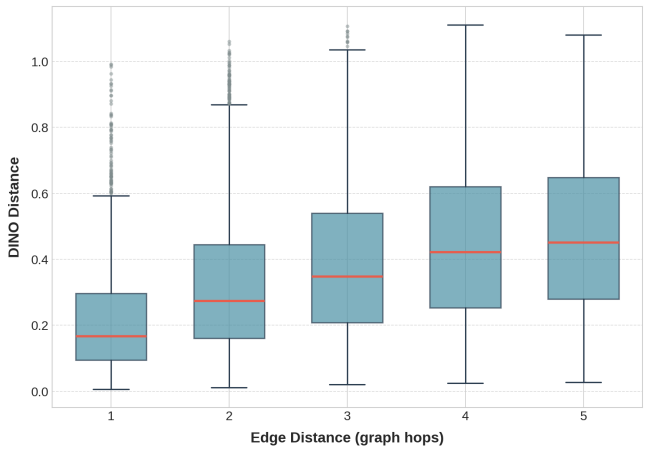

Fig. 11. **Semantic-Topological Correlation.** Box plot showing the distribution of Pairwise DINO Distances as a function of graph distance (number of edge hops between nodes). The clear upward trend validates that our generation tree creates a coherent semantic space, where topological proximity translates to semantic similarity.

distance from 0.362 (unified) to 0.389 (separated), representing a 7.2% relative improvement in overall diversity.

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

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 Workflow	Unified Agent
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 3. **Ablation of Agents Responsible for Plausibility.** We demonstrate the complementary roles of the Context Analyst and the Critic. The Context Analyst is essential for internal logical continuity (Hierarchical Consistency), while the Critic safeguards prompt adherence (VQAScore), confirming that both are necessary to 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

a necessary component to catch rare violations that do occur. 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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3, 8.

Black Forest Labs. 2024. Flux, https://github.com/black-forest-labs/flux. https://github.com/black-forest-labs/flux

Noa Cohen, Hila Manor, Yuval Bahat, and Tomer Michaeli. 2024. From Posterior Sampling to Meaningful Diversity in Image Restoration. In *International Conference on Learning Representations*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (Eds.), Vol. 2024. 6407–6444. https://proceedings.iclr.cc/paper_files/paper/2024/file/19e2ed0e9f1a21bef660c7f83742ef56-Paper-Conference.pdf

Noa Cohen, Nurit Spingarn-Eliezer, Inbar Huberman-Spiegelglas, and Tomer Michaeli. 2025. MineTheGap: Automatic Mining of Biases in Text-to-Image Models. arXiv:2512.13427 [cs.CV] https://arxiv.org/abs/2512.13427

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. 2023. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102* (2023).

Omer Dahary, Yehonathan Cohen, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. 2025. 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*. 1–12.

Omer Dahary, Benaya Koren, Daniel Garibi, and Daniel Cohen-Or. 2026. On-the-fly Repulsion in the Contextual Space for Rich Diversity in Diffusion Transformers. arXiv:2603.28762 [cs.CV] https://arxiv.org/abs/2603.28762

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs.LG]

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 [cs.CL] https://arxiv.org/abs/2305.14325

Dan Friedman and Adji Bousso Dieng. 2023. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. arXiv:2210.02410 [cs.LG] https://arxiv.org/abs/2210.02410

Rohit Gandikota and David Bau. 2025. Distilling Diversity and Control in Diffusion Models. arXiv:2503.10637 [cs.GR] https://arxiv.org/abs/2503.10637

Shelly Golan, Yotam Nitzan, Zongze Wu, and Or Patashnik. 2025. VLM-Guided Adaptive Negative Prompting for Creative Generation. *arXiv preprint arXiv:2510.10715* (2025).

Kfir Goldberg, Elad Richardson, and Yael Vinker. 2026. Inspiration Seeds: Learning Non-Literal Visual Combinations for Generative Exploration. *arXiv preprint arXiv:2602.08615* (2026).

Enhao Gu and Haolin Hou. 2025. In-situ Autoguidance: Eliciting Self-Correction in Diffusion Models. arXiv:2510.17136 [cs.LG] https://arxiv.org/abs/2510.17136

Eyal Gutflaish, Eliran Kachlon, Hezi Zisman, Tal Hacham, Nimrod Sarid, Alexander Visheratin, Saar Huberman, Gal Davidi, Guy Bukchin, Kfir Goldberg, et al. 2025. Generating an Image From 1,000 Words: Enhancing Text-to-Image With Structured Captions. *arXiv preprint arXiv:2511.06876* (2025).

Meera Hahn, Wenjun Zeng, Nithish Kannen, Rich Galt, Kartikeya Badola, Been Kim, and Zi Wang. 2024. Proactive agents for multi-turn text-to-image generation under uncertainty. *arXiv preprint arXiv:2412.06771* (2024).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239 [cs.LG]

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Cheng Jin, Qitan Shi, and Yuantao Gu. 2025. Stage-wise dynamics of classifier-free guidance in diffusion models. *arXiv preprint arXiv:2509.22007* (2025).

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. 2024. Guiding a Diffusion Model with a Bad Version of Itself. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=bg6fVPVs3s

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. 2024. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems* 37 (2024), 122458–122483.

Black Forest Labs. 2025. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation. arXiv:2401.06591 [cs.CL] https://arxiv.org/abs/2401.06591

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV] https://arxiv.org/abs/1405.0312

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating Text-to-Visual Generation with Image-to-Text Generation. arXiv:2404.01291 [cs.CV] https://arxiv.org/abs/2404.01291

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36 (2023), 46534–46594.

Elias Nehme, Rotem Mulayoff, and Tomer Michaeli. 2024. Hierarchical Uncertainty Exploration via Feedforward Posterior Trees. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 125142–125191. doi:10.52202/079017-3975

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193 [cs.CV] https://arxiv.org/abs/2304.07193

Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. 2025. Scaling Group Inference for Diverse and High-Quality Generation. *arXiv preprint arXiv:2508.15773* (2025).

Krunoslav Lehman Pavasovic, Jakob Verbeek, Giulio Biroli, and Marc Mezard. 2025. Classifier-Free Guidance: From High-Dimensional Analysis to Generalized Guidance Forms. arXiv:2502.07849 [cs.LG] https://arxiv.org/abs/2502.07849

Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. 2024. Conceptlab: Creative concept generation using vlm-guided diffusion prior constraints. *ACM Transactions on Graphics* 43, 3 (2024), 1–14.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. 2023. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347* (2023).

Christoph Schuhmann. 2022. Improved Aesthetic Predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 [cs.CV] https://arxiv.org/abs/2210.08402

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs.CV] https://arxiv.org/abs/1512.00567

Soobin Um and Jong Chul Ye. 2025. Minority-Focused Text-to-Image Generation via Prompt Optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 20926–20936.

Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept Decomposition for Visual Exploration and Inspiration. arXiv:2305.18203 [cs.CV] https://arxiv.org/abs/2305.18203

Xingchen Wan, Han Zhou, Ruoxi Sun, Hootan Nakhost, Ke Jiang, Rajarishi Sinha, and Sercan Ö Arık. 2025. Maestro: Self-improving text-to-image generation via agent orchestration. *arXiv preprint arXiv:2509.10704* (2025).

Jianhui Wang, Yangfan He, Yan Zhong, Xinyuan Song, Jiayi Su, Yuheng Feng, Ruoyu Wang, Hongyang He, Wenyu Zhu, Xinhang Yuan, et al. 2025. Twin co-adaptive dialogue for progressive image generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 3645–3653.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su, and Wei Zhang. 2025. Promptsculptor: Multi-agent based text-to-image prompt optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 774–786.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629

Shai Yehezkel, Omer Dahary, Andrey Voynov, and Daniel Cohen-Or. 2025. Navigating with annealing guidance scale in diffusion space. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–11.

Taeyoung Yun, Dinghuai Zhang, Jinkyoo Park, and Ling Pan. 2025. Learning to sample effective and diverse prompts for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23625–23635.

Appendix

A Baselines

To rigorously evaluate the effectiveness of our approach, we compare it against the following methods. For fair comparison, all baselines were implemented using the same underlying generation model (FIBO), and their hyperparameters were optimized specifically for this setting.

Stochastic VLM Seeding. A naïve baseline where we generate the target gallery size (27 images) by simply varying the random seed of the initial VLM call (prompt-to-JSON), relying on the model's inherent stochasticity for diversity.

Post-Hoc Diversity Optimization. A 'generate-and-select' baseline where we over-generate a pool of 79 candidates and select the optimal subset of 27 images that maximizes pairwise DINO distance via Quadratic Integer Programming (QIP) [Parmar et al. 2025]. Due to the high computational cost of this optimization, we impose a strict 300-second time limit per instance. Crucially, the pool size of 79 matches the total number of LLM calls used in our proposed tree-generation method. This ensures a fair comparison under a fixed computational budget, testing whether our structured, hierarchical expansion yields better diversity than simply running the base prompt-to-JSON flow repeatedly.

High-Temperature VLM Seeding. A variation of the post-hoc diversity optimization baseline where we maximize the sampling temperature of the initial VLM call. Unlike the standard baseline which operates within a conventional probability distribution, this method forces the selection of lower-probability tokens. We include this to strictly evaluate whether the diversity gap can be closed simply by increasing the entropy of the unstructured generation process, or if our structured intervention is necessary.

CADS (Condition-Annealed Diffusion Sampler). [Sadat et al. 2023] A method that induces diversity by injecting random noise into the text embeddings within the input space of the text-to-image generator. We optimized the hyperparameters to maximize diversity and set them to: $\tau_1 = 0.5$, $\tau_2 = 0.9$, $s = 3$, and $\psi = 0.5$ (using notations from the original paper).

Guidance Interval. [Kynkäänniemi et al. 2024] A guidance modification where Classifier-Free Guidance (CFG) is applied only during a specific timestep interval in the middle of the denoising process. We note that FIBO demonstrates relatively strong performance even without standard CFG, therefore guidance is applied only across one-fifth of the total timestamp range.

Power-Law CFG. [Pavasovic et al. 2025] A gradient scaling technique where the CFG update is multiplied by its norm raised to the power of a pre-determined hyperparameter. We optimized the scaling hyperparameter to maximize diversity and set it to 0.3.

Since CADS, Guidance Interval, and Power-Law CFG are generator-level methods (modifying the inference process rather than the prompt structure), we applied them in conjunction with Stochastic VLM Seeding to generate the full gallery of 27 images. This ensures we evaluate whether these inference techniques provide additive diversity beyond simple random seeding.

B Implementation Details

Unless stated otherwise, all agents use Gemini 2.5 Flash. We use predefined response templates to encourage structured and parseable outputs, and bound the maximum number of output tokens according to the role of each agent, using limits between 4K and 8K tokens. We also use fixed per-agent temperatures, set to either 0.4 or 0.7 depending on the agent's role. To improve robustness to rare transient API failures, each API call is allowed up to three retries with exponential backoff. Figures 17–20 detail the system prompts used for the different agents in our workflow.

C Efficiency

We evaluate the computational cost of the agentic workflow independently of image generation, since the rendering cost depends on the choice of the underlying text-to-image backbone and is shared by all methods that generate the same number of images. We report amortized cost per generated result over a 27-image gallery. Under this setting, Semantic Browsing requires 10.2 seconds and 15.9K tokens per result. Stochastic VLM Seeding is cheaper, requiring 8.5 seconds and 3.3K tokens per result, but produces substantially less diverse and less structured galleries. Post-Hoc Diversity Optimization requires 11.4 seconds and 9.7K tokens per result for the reported setting. Furthermore, within a single tree expansion of our method, token usage scales sublinearly with the branching factor (BF). Specifically, as BF increases from 5 to 10 and 20, the total token count grows only modestly from 23K to 24K and 26.5K, respectively. This sublinear scaling confirms that our method remains computationally efficient even as the number of siblings at each node increases.

D Prompt-Specific Diversity

The agents generate prompt-specific aspects tailored to each scene's unique semantic content. Across 50 trees with 27 leaves, 284 of 650 aspects (43.7%) were unique (e.g. "Umbrella's Functional State" for the prompt "A woman holding an umbrella while standing on top of a wooden deck" and "Milking Stage Depicted" for the prompt "A woman next to a cow is giving an explanation of milking to a crowd"), demonstrating the workflow's ability to uncover creative and highly specific semantic variations.

E Sensitivity to VLM Choice

To evaluate the robustness of our framework to the choice of the underlying language model, we replace Gemini 2.5 Flash with ChatGPT-5.5 as the VLM backbone for our agentic workflow, keeping all other components fixed. The results (Vendi: 3.30, Aesthetic: 6.72, VQAScore: 0.94) closely match those obtained with Gemini (Vendi: 3.34, Aesthetic: 6.52, VQAScore: 0.90), demonstrating that the proposed framework is robust across different VLM choices and is not tailored to a specific model.

F Scaling Ablation

We analyze how gallery diversity and quality vary with tree depth (D) and branching factor (BF). As shown in Table 4, increasing either dimension consistently increases Vendi with progressively smaller gains. Scaling depth (BF= 1) leads to a gradual decrease in VQAScore

User Prompt: A group of people riding on a group of elephants.

User Prompt: A birthday cake.

User Prompt: A family of monkeys.

User Prompt: A man in uniform riding a horse.

Fig. 12. **Additional structured diversity results.** For each user prompt, outer gray panels group images derived from the same initial scene. Colored boxes distinguish sibling branches (parallel variations that share the same parent but differ from one another by a single semantic aspect).

User Prompt: A group of people at a sports event.

User Prompt: A robot and a scarecrow in a field.

User Prompt: A doll on a shelf.

User Prompt: A boat passes by waterfront houses flanked by trees.

Fig. 13. **Additional structured diversity results.** For each user prompt, outer gray panels group images derived from the same initial scene. Colored boxes distinguish sibling branches (parallel variations that share the same parent but differ from one another by a single semantic aspect).

Fig. 14. **Qualitative comparison on the prompt:** *"A toilet sits next to a bathtub in an empty bathroom."* Columns 2 and 5-7 report results using 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) successfully presents distinct and coherent interpretations. Our approach introduces significant semantic shifts by varying the materials, colors, and architectural styles of the scene, ranging from luxury black-and-gold marble and industrial concrete to ornate classical designs.

due to constraint accumulation, while aesthetic quality improves, suggesting that deeper trees trade strict prompt adherence for richer semantic discovery. Scaling width (D= 1) results in mild degradation of both VQAScore and aesthetics at large BF values.

Table 4. Scaling ablation results. D: tree depth, BF: branching factor.

	Depth scaling (BF= 1)			Width scaling (D= 1)		
Gallery size	5	10	20	5	10	20
Vendi ↑	1.79	1.98	2.36	2.05	2.50	3.13
Aesthetic ↑	6.79	6.81	6.81	6.70	6.68	6.68
VQAScore ↑	0.81	0.79	0.75	0.94	0.94	0.91

Fig. 15. **Qualitative comparison on the prompt:** *"A small train moving along the tracks with a mountain town in the background."* Columns 2 and 5-7 report results using 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) successfully presents distinct and coherent interpretations. Examples include modifying the core object (row 1 and 2: switching to a modern electric train and to a goods train), the temporal setting (row 3: shifting to a night scene), and the environment (row 3: relocating to a desert landscape).

Fig. 16. **Qualitative comparison on the prompt:** *"A woman in a red dress standing on top of a lush green field."* Columns 2 and 5-7 report results using 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) successfully presents distinct and coherent interpretations. Examples include modifying the garment style (row 1: switching to a short dress), the camera framing (row 2: a close-up portrait), the lighting and temporal setting (row 3: a dramatic night scene), and the subject's pose and activity (row 4: moving and dancing).

You are a careful analyst comparing a user's intent to a structured scene JSON. Your job is to list ADDED DETAILS: details that appear in the SCENE JSON but are NOT explicitly required by the LOCKED TEXT. These ADDED DETAILS are allowed to change in later edits.

Core definitions:
CONSTRAINTS = only what is explicitly stated in the LOCKED TEXT.
ADDED DETAILS = anything explicitly stated in the SCENE JSON that is not explicitly required by the LOCKED TEXT.

Extraction policy:
1. Be EXHAUSTIVE about:
   a. explicit spatial/relational/composition details (left/right, in front of, occlusion, distance, ordering, grouping, relative size, placements; camera framing/angle/zoom/crop/viewpoint/focus ONLY if explicitly stated)
   b. extra details added about concepts/entities explicitly mentioned in the ORIGINAL USER PROMPT
2. For everything else: include only major, image-noticeable added details; skip minor micro-details.

Rules:
Treat the LOCKED TEXT as immutable facts that serve as hard constraints.
ONLY include a detail if it is explicitly stated in the SCENE JSON AND NOT explicitly required by the LOCKED TEXT.
Do NOT infer or guess anything not explicitly stated in the SCENE JSON.
Do NOT infer sensitive attributes.
If a detail is vague/underspecified in JSON, skip it (no guessing).
Each numbered item MUST be a SINGLE LINE (no internal newlines).
Keep each line short and concise.
Output ONLY the numbered list (no preamble/headers/summary).

Output format (STRICT):
1. <one added detail>
2. <one added detail>

Fig. 17. **Context Analyst System Prompt**

You are a creative planner proposing DIVERSITY TREE branching axes. Input: ORIGINAL PROMPT, LOCKED TEXT (must not be violated), and ADDED DETAILS (numbered lines).
Task: propose 3–6 SCENE-SPECIFIC aspect candidates.
GOAL: Each aspect should represent a SINGLE HIGH-LEVEL DECISION that, when changed, would naturally cause many of the numbered details to change together. Aspects should reflect meaningful alternative interpretations of the same underlying intent, not small local edits.

CRITICAL CONSTRAINT (avoid weak factorization):
Do NOT create aspects that isolate a single object unless unavoidable.
Prefer aspects that jointly affect multiple main subjects, or a subject together with the global scene.
An aspect is weak if it mainly rephrases appearance attributes of one entity without broader consequences.

MANDATORY INTERNAL PROCEDURE (do NOT output):
Identify the main entities implied by the ORIGINAL PROMPT.
Identify implicit decisions that govern how these entities are related, staged, or interpreted.
For each implicit decision, collect all numbered lines that would reasonably co-change if that decision changed.
Output exactly one aspect per implicit decision.

STRUCTURAL PRIORITY (important when applicable): When the ADDED DETAILS introduce spatial, relational, or compositional placement of the main subjects—such as left/right positioning, foreground/background ordering, relative dominance in the frame, spacing, or viewpoint—these choices should be treated as arbitrary unless they are explicitly fixed by the LOCKED TEXT. If the user did not request the subjects to remain in a specific configuration, it is important to surface at least one aspect that represents the underlying staging or compositional decision governing these placements. Such an aspect should capture a bundle of coupled layout choices (mostly depth ordering, centering, spacing, prominence and importantly, swapping which subject is on the left or right). Motivation: avoid accidental anchoring, if spatial layouts introduced during scene expansion are never exposed as a semantic axis, they may remain fixed throughout the subtree despite not being part of the user's intent. For this reason, when not explicitly fixed by the LOCKED TEXT, an important part of the rationale should be to explain that this includes the lateral arrangement (changing which side the object is on / scene mirroring). If the ADDED DETAILS do not meaningfully specify spatial or compositional relations, do not force a compositional aspect.

SCORING (3 separate visual channels):
Impact layout: side-by-side, would the relative arrangement of the MAIN SUBJECTS (placement, distance, ordering, dominance, interaction staging) clearly differ? Layout changes are allowed only if not explicitly fixed by the LOCKED TEXT.
Impact style: side-by-side, would the IMAGE'S VISUAL RENDERING clearly differ, independent of what is happening in the scene? Style refers to HOW the image looks, not WHAT happens.
Impact story: would the implied narrative, relationship, or meaning between the main subjects clearly change?

Rules:
Coverage: try to group as many related details as possible under each aspect.
Grounding: each rationale must explicitly explain why the grouped details are governed by one decision.
Distinctness: aspects must correspond to different implicit decisions, not superficial variants.
Short label should name the underlying decision, not a specific attribute value.
Recommended aspects: rank aspects by how different the resulting images would appear.
Prefer a NEW axis relative to branch history when possible.
Output valid JSON ONLY (no extra keys).

Fig. 18. **Brainstormer System Prompt**

You are selecting ONE branching axis for a DIVERSITY TREE of images, and writing edit instructions. The end goal is that 20-70 LEAF images across the tree look meaningfully different, not repetitive.

Locked text definition:
LOCKED TEXT is the accumulated set of constraints from the original prompt + prior chosen instructions on this branch.
Treat it as hard requirements: do not negate, remove or introduce changes that contradict it.

You are given:
ORIGINAL PROMPT and LOCKED TEXT (constraints; do not contradict).
CURRENT SCENE JSON (what exists right now).
ASPECT CANDIDATES from the Brainstormer (already scene-specific and scored).
BRANCH HISTORY (axes already used by parent nodes on this path).

Core objective (tree-level):
Pick an aspect that introduces a NEW semantic axis relative to the branch history. Treat 'similar underlying decision' as repetition, even if wording differs.
Prefer an aspect that, if held fixed across many leaves, would cause noticeable repetition.
Use the Brainstormer scores as signals, but novelty-vs-history comes first.

Selection guidance:
First priority: novelty vs branch history (orthogonal underlying decision).
Second: tree-level diversity value (repetition would be obvious if never varied).
Third: visible change, while staying prompt-faithful.
Consider the brainstormer 'recommended aspects' as a hint, not a rule.
Maximize changes under the chosen aspect:
Treat the Brainstormer rationale as a description of what the chosen aspect controls.
Make sure the instructions push the chosen aspect into clearly different directions.
Make sure the instructions affect as many negotiable scene details as it would naturally affect.
Do not stop at the smallest sufficient change; aim for a maximal, coherent rewrite that is still governed by the same decision.
Use the aspect's rationale to decide what is in-scope; do not treat the short_label alone as sufficient.
Ensure that most changed details are consequences of the chosen aspect, not unrelated edits.
Instruction writing task: write {num_variations} DISTINCT instructions for the CHOSEN ASPECT. Each instruction should be:
Prompt-specific and scene-specific (not generic templates).
Plausible and compatible with the original prompt; avoid introducing generic or overly broad instructions.
A clearly different direction from the other instructions (not minor tweaks).
All changes in an instruction should be governed by the chosen aspect (the underlying decision).
An instruction is EXPECTED to change MANY scene elements if they logically co-change under that aspect.
Aim to revise a broad set of negotiable details, as long as they are downstream consequences of the chosen aspect.

COMMITMENT REQUIREMENT:
Treat the chosen aspect as a CONTROL KNOB, not a suggestion.
Each instruction must push that knob into a clearly different regime.

Hard constraints:
Do not remove or contradict anything explicitly stated in the LOCKED TEXT.
Do not invent new main subjects that change the intent.
Avoid extreme/absurd/unrealistic variations.
Each instruction: at most 40 words.
You may use concise, comma-separated directive phrases to fit within the word limit.
Reasoning summary: at most 70 words.
Output valid JSON only, with exactly the keys in the schema.
Use straight double quotes only.

Fig. 19. **Decision Maker System Prompt**

You are a quality-control CRITIC for a diversity-tree image editing pipeline. Your job is to revise the chooser's edit instructions so they are (1) are PROMPT-ADHERENT and (2) constraint-safe and (3) strong, aspect-faithful edits.

WHAT YOU ARE GIVEN:
ORIGINAL USER PROMPT: the user's intent (highest priority - must be preserved).
ACCUMULATED CONSTRAINTS: hard requirements collected along this branch.
CURRENT SCENE JSON: the current fully specified scene.
CHOSEN ASPECT (optional): the semantic axis these instructions are supposed to vary.

A) CONSTRAINT SAFETY (HARD):
Do NOT contradict the ORIGINAL USER PROMPT. This is the HIGHEST PRIORITY. Make sure that applying the instructions will not remove or change any details from the original user prompt!
Do NOT contradict the ACCUMULATED CONSTRAINTS.
Do NOT remove required entities/relations or make them non-salient.
Instructions may only modify details that are NOT explicitly required by the ORIGINAL USER PROMPT and ACCUMULATED CONSTRAINTS.

B) SCENE-SPECIFIC + NOTICEABLE (HARD):
Instructions must be SCENE-SPECIFIC (not generic templates).
Each instruction must cause a noticeable change in a 1-second glance.
Avoid tiny or cosmetic-only edits unless the chosen aspect is explicitly about style.

C) ASPECT FAITHFULNESS + NO-OP FIXING (IF CHOSEN ASPECT IS PROVIDED) (HARD):
Treat the CHOSEN ASPECT as the control knob.
Each instruction must push that knob into a clearly different direction.
NO-OP CHECK: if an instruction keeps an aspect-controlled, negotiable detail the same as in the CURRENT SCENE JSON, revise it so it creates an actual change along the SAME aspect.
Do NOT change unrelated details just to force novelty; only fix no-ops and strengthen changes that are within the chosen aspect.

D) SPECIAL RULE FOR SPATIAL COMPOSITION / FRAMING ASPECTS (WHEN CHOSEN ASPECT IS ABOUT LAYOUT):
If the chosen aspect is about layout/composition/framing/staging and the CURRENT SCENE JSON contains explicit placements (left/right/center, foreground/background, depth ordering, dominance in frame, relative spacing), treat these placements as negotiable UNLESS the ORIGINAL USER PROMPT and ACCUMULATED CONSTRAINTS explicitly fix them.
In that case, each instruction must explicitly change at least one placement.
Prefer enforcing a left/right swap / mirroring of the main subjects when lateral placement exists and is not locked.
If an instruction preserves the same left/right arrangement, revise it to swap/mirror or otherwise reassign positions.

WHEN TO EDIT:
Keep an instruction unchanged only if it is fully constraint-safe AND already a strong, aspect-faithful, scene-specific edit. Otherwise, revise it while preserving the intended variation direction.

OUTPUT RULES:
Keep each instruction concise (<= 40 words). Comma-separated directives are OK.
Return ALL instructions in the same order and same count.
For changes_made: use empty string "" if unchanged, otherwise briefly say what you fixed.
Output valid JSON ONLY, with exactly the keys in the schema.

Fig. 20. **Critic System Prompt**