coDrawAgents: A Multi-Agent Dialogue Framework for Compositional Image Generation

Supplementary Material

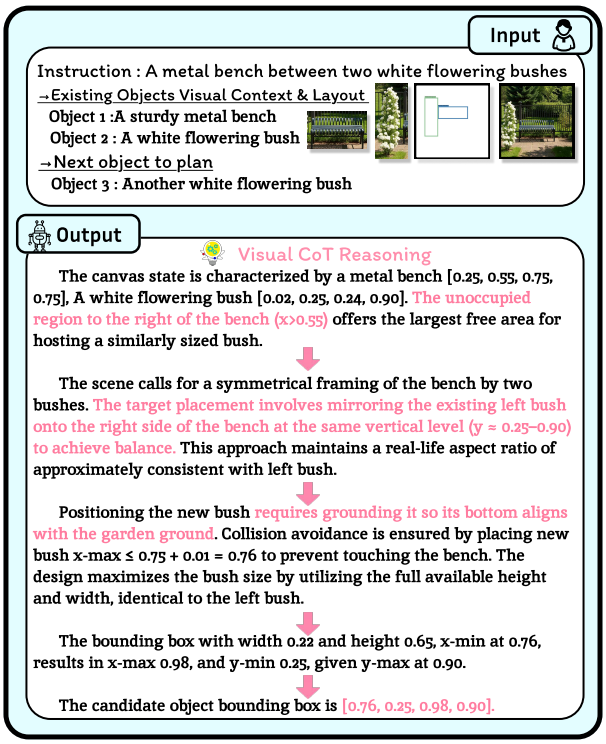


Figure 1. A reasoning example of the VS-COT module in our framework

1. Details of Visual CoT (VCoT) Reasoning

Figure 1 illustrates the detailed reasoning steps of our Visual Spatial CoT (VCoT) module. Given the current canvas state and instruction, VCoT performs the following steps:

1. **Scene parsing.** The module first parses the existing layout, identifying objects and their bounding boxes. For example, a metal bench is located at [0.25, 0.55, 0.75, 0.75], and a white flowering bush is positioned at [0.02, 0.25, 0.34, 0.90].
2. **Free-space identification.** VCoT then analyzes the unoccupied regions of the canvas. In this case, the right side of the bench (x $\approx$ 0.25–0.95) is identified as the feasible area for placing another bush.
3. **Symmetry reasoning.** To maintain balanced composition, the target placement is chosen to mirror the existing bush on the left, aligning along the same vertical level (y $\approx$ 0.90).
4. **Grounding and constraint enforcement.** The candidate bounding box is grounded to the garden floor (bottom aligned), ensuring physical plausibility. Collision constraints are checked to avoid overlap with the bench by setting the bush's x-max $\leq$ 0.76.
5. **Bounding box refinement.** The box dimensions are adjusted to maximize use of available space while keeping consistency with the left bush. The final bounding box is given as [0.76, 0.25, 0.98, 0.90].

Through these step-by-step spatial reasoning processes, VCoT generates placements that are physically valid, compositionally balanced, and text-aligned.

2. Details of Prompts

Figure 2 presents the detailed design of our multi-agent prompting strategy.

3. Details of Multi-agent collaboration

At each step, 3DIS first predicts a depth map and then renders the corresponding image using the depth-to-image model. During multi-agent involvement, we iteratively accumulate the depth information from the previous iteration, estimate a refined depth map for the next iteration, and synthesize an updated image based on this new depth. At each iteration, the intermediate image serves only as planning context for the subsequent step, enabling the system to refine local regions while keeping the global structure stable.

Figure 3 illustrates the details of our multi-agent collaboration framework. Given an input prompt, the Interpreter determines the generation mode and decomposes the prompt into instance descriptions. The Planner then proposes an initial layout, which is further checked and refined by the Checkers. Using the refined layouts, the Painter performs iterative planning and synthesis, where each intermediate painting—along with its grounding information—serves as visual context for the next iteration. Finally, the Painter produces the final image.

4. Additional results on DPG Bench

Figure 5, 6 and 7 presents six representative cases from DPG-Bench, covering diverse everyday and imaginative scenarios. The prompts feature multiple objects, long textual descriptions, and complex interactions, such as indoor arrangements (books, couches, kitchen scenes), dynamic activities (a surfer riding waves, a man and dog playing with a frisbee), and even global landmarks (Sydney Opera House, Eiffel Tower, Mount Everest). Across all these

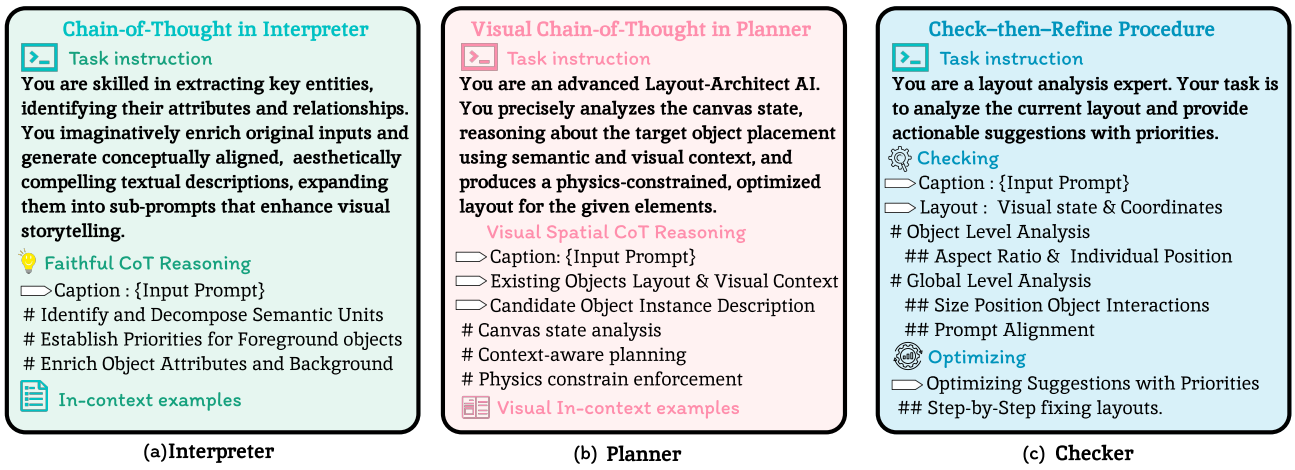


Figure 2. Prompting structure for agents in our framework.

cases, coDrawAgents produces images that remain highly faithful to the text, accurately capturing object positions, counts, and attributes while preserving visual coherence and style. These results further confirm that our multi-agent framework scales effectively to the challenging long-context, multi-object compositions posed by DPG-Bench.

Notably, the inclusion of layout-free examples, where instance decomposition and bounding boxes are not available, demonstrates the flexibility of our approach, which can adaptively select between layout-free and layout-aware modes to accommodate varying text prompt requirements.

5. Additional results on GenEval

Figures 8 present our generative results on GenEval. coDrawAgents accurately captures object attributes and spatial relationships across diverse scenarios. The framework successfully generates realistic and coherent images, such as a dog positioned to the right of a teddy bear, as well as indoor scenes involving a brown dining table and a white sofa frame.

Notably, coDrawAgents exhibits strong generalization capabilities in handling unconventional or semantically implausible compositions, such as a majestic brown horse placed alongside a leather couch or a computer keyboard, which are rare or unrealistic in real-world contexts. These cases highlight the model's robustness in following uncommon or imaginative prompts.

6. Limitation

While coDrawAgents Dialogue demonstrates significant progress in compositional text-to-image generation, it still has several limitations:

First, the multi-agent system, while beneficial for quality and compositional accuracy, introduces a computational overhead. The framework requires more processing time compared to single-pass methods due to the multi-agent calls. However, empirical analysis shows coDrawAgents Dialogue still achieves competitive inference efficiency, outperforming many existing methods despite its iterative nature. Further optimization of the multi-agent loop remains a key area for future work.

Second, the performance of our Painter is inherently dependent on the underlying T2I and L2I models. This dependency means that limitations of the base models, such as imperfect attribute rendering or biased visual priors, may propagate into coDrawAgents, e.g., "a radish with black skin". Conversely, it also indicates that coDrawAgents will naturally benefit from future advances in text-to-image and layout-to-image generation.

Third, the Planner and Checker rely on multimodal LLMs for layout reasoning and error detection, making the system susceptible to LLM-specific issues such as hallucination and overconfidence in incorrect layouts. These limitations may lead to invalid object placements or missed corrections, especially in highly compositional or ambiguous prompts. Conversely, it also indicates that coDrawAgents will naturally benefit from future advances in more reliable and grounded LLMs with reduced hallucination tendencies.

Finally, as with most iterative frameworks, coDrawAgents may be affected by error accumulation across iterations. For instance, small placement inaccuracies in early steps can propagate if not fully corrected by the Checker. Nevertheless, our design explicitly mitigates this risk by introducing verification and refinement mechanism, and we observe that the overall error accumulation is significantly lower than in single-pass generation pipelines

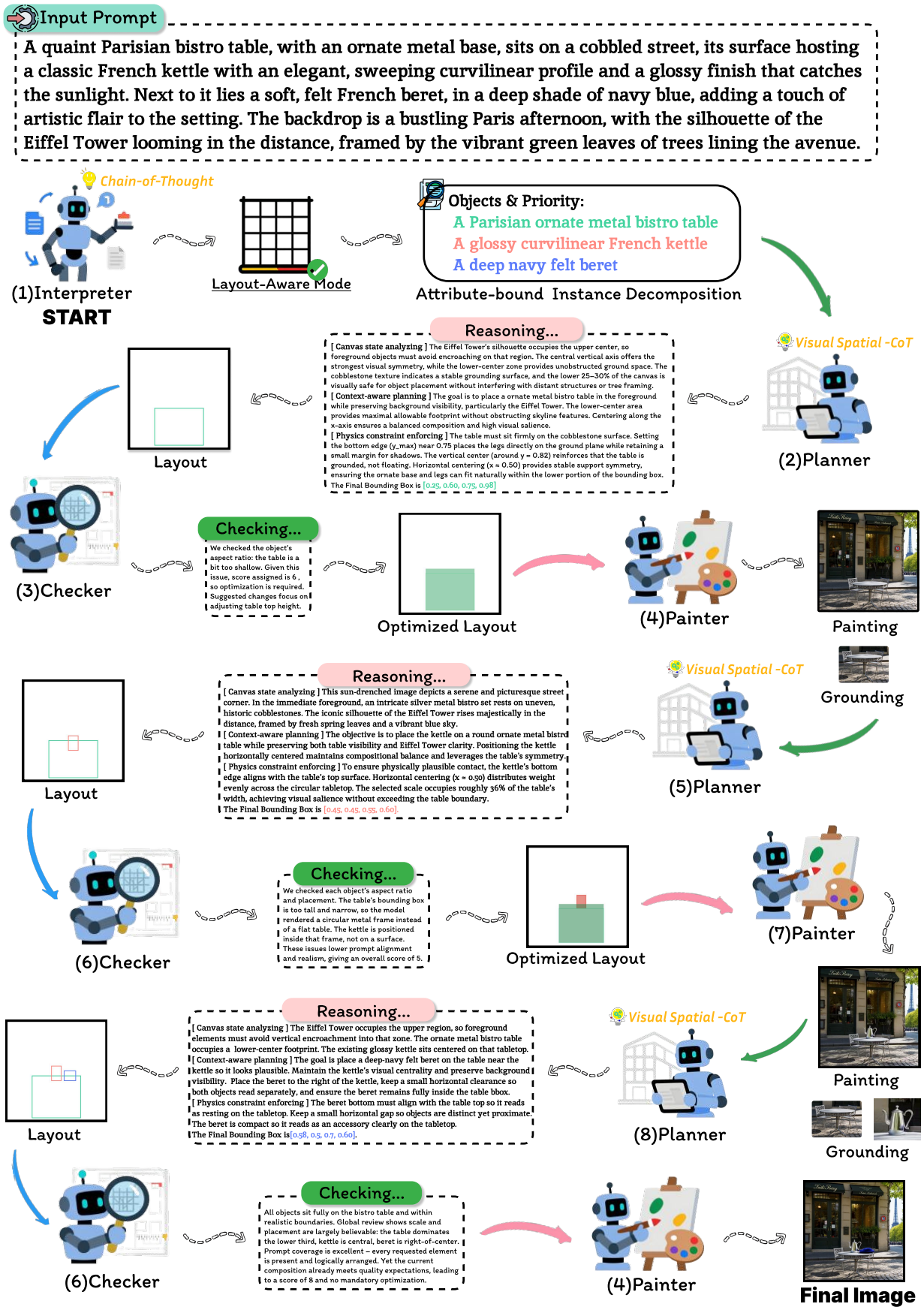


Figure 3. Multi-agent collaboration.

7. Qualitative Results

View of kitchen and study area, kitchen with tiled countertop and floor, oven and overhead microwave, spice rack, kettle, and in study area, wood floor, shelves, window, desk with desk lamp, and a padded, rolling office chair.

a whimsical illustration of a small, white baby daikon radish with rosy cheeks and green shoots atop its head, donning a pink, frilly tutu. It is walking a brown, fluffy dog on a red leash, which looks up at the radish with a playful expression. The background features a simple, pastel-colored path that winds through a grassy field.

a large, round orange pumpkin carved with a smiling face, sitting on a wooden table. Inside the hollowed-out pumpkin, a small flickering candle casts a warm glow through the cut-out eyes and mouth. The pumpkin is surrounded by a scattering of fallen autumn leaves.

An imaginative scene where the iconic Sydney Opera House, with its white sail-like shells, sits prominently on the left. To the right, the Eiffel Tower, constructed of intricate iron lattice work, towers over the landscape. Behind both landmarks, the majestic Mount Everest looms, its snow-capped peak piercing the sky.

An abstract oil painting that depicts a chaotic blend of vibrant colors and swirling patterns, giving the impression of a vast, disorienting landscape. The canvas is filled with bold strokes of reds, blues, and yellows that seem to clash and compete for space, symbolizing the complexity and confusion of navigating through life. Amidst the turmoil, a small, indistinct figure appears to be wandering, searching for direction in the overwhelming expanse.

A colorful collection of four cartoon-styled calendars, each uniquely illustrating the essence of a different season. The spring calendar bursts with shades of green and pink, featuring blooming flowers and sprouting leaves. The summer calendar glows with vibrant sun motifs and vivid blue skies. Autumn is represented with warm oranges and browns, showcasing falling leaves and harvest themes. The winter calendar is adorned with soft whites and blues, depicting snowy scenes and cozy fireside images. Each calendar is distinct, yet they all share a whimsical charm that captures the spirit of their respective seasons.

A picturesque outdoor scene featuring a ceramic vase prominently placed to the left of a lush, green lawn. The vase, with its smooth texture and intricate patterns, stands in the foreground, with the expansive, clear blue sky stretching overhead. Beyond the vase, a wooden bench can be seen, slightly obscured by the vase's presence. To the right, a dense, leafy bush rises up against the sky, situated just above a paved street that runs adjacent to the bush.

A playful collection of 2x2 emoji icons, each resembling a vibrant macaron with a distinct facial expression. The top left macaron is a sunny yellow with a beaming smile, while the top right is a fiery red with furrowed brows and an angry scowl. Below them, the bottom left is a bright blue with wide, surprised eyes, and the bottom right is a soft lavender with a tearful, sobbing face. Each of the macaron emojis is whimsically topped with a miniature brown cowboy hat, adding a touch of whimsy to their appearance.

a modern storefront with large glass windows and a bold sign above the entrance that reads 'openai' in sleek, white lettering. The facade is painted in a muted gray, complementing the contemporary design. Inside, through the transparent windows, one can see rows of neatly arranged products and a few customers browsing.

a brightly colored storefront with large, bold letters spelling out 'AwesomePurchase' above the entrance. The shop's window displays are neatly arranged with an array of products, and a small, potted plant sits to the left of the door. The facade of the building is a clean, modern white, contrasting with the vibrant signage.

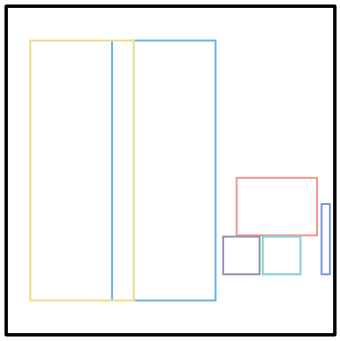
A collection of nine vibrant VIP sign icons, indicating membership or exclusive status. Each label varies in color, providing a rainbow spectrum from red to violet, and they are designed with a sleek, glossy texture. The symbols feature a simple, bold font that stands out against the solid background, rendering them ideal for vector illustrations where distinction is key.

In the distance, towering black mountains with their peaks blanketed in thick layers of snow stand majestically. Against this dramatic backdrop, a flock of black birds is captured in their dynamic mid-flight, crisscrossing the scene with elegance and energy. Above them, the sky is a tapestry of deep grays clashing with the remnants of serene blue, creating a striking contrast that defines the horizon.

Figure 4. Qualitative results.

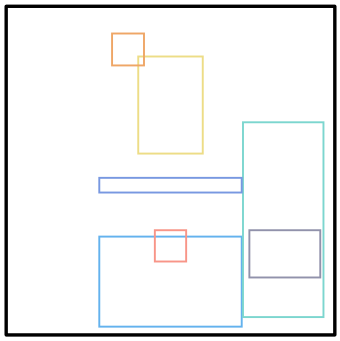
Instance descriptions | **Layout** | **Image**

A man with a relaxed posture
A woman in a bland kitchen
A modern kitchen microwave
A plate with food
An empty kitchen plate
A wine bottle



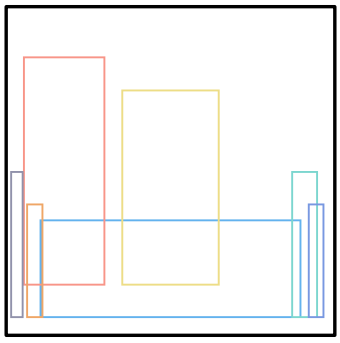
A man, his arm across the woman next to him, stands in a blandly colored kitchen area, in front of a black-rimmed window, next to a counter with a microwave, plates, with and without food, and wine bottles.

A wooden vanity
A bathroom mirror
An integrated bathroom sink
A tiled shower
A molded plastic bathtub
A small practical shelf
A small bathroom lamp



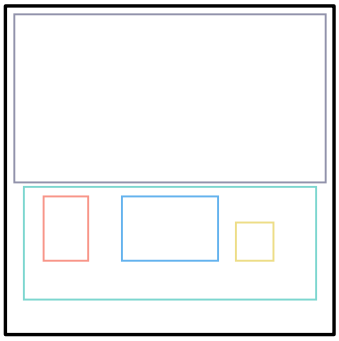
a photo of a brown dining table and a white suitcase

A contemporary sectional sofa
A large decorative mirror
A filled bookshelf
A modern floor lamp
A modern floor lamp
A vase of fresh flowers
A vase of fresh flowers

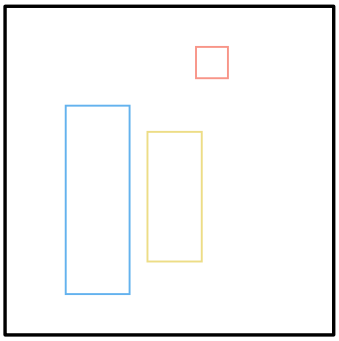


A sectional sofa in a front room with a bookshelf and mirror with two floor lamps and two vases of flowers on either side。

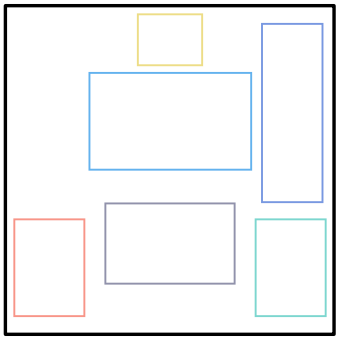
Figure 5. Qualitative results on DPG.

Instance descriptions	Layout	Image
A stack of blue books A red apple A bell A wooden teacher's desk A dusty old chalkboard		

A dusty old chalkboard fills the background of this image, which features a wooden teacher's desk that has blue books, a red apple, and a bell placed on top.

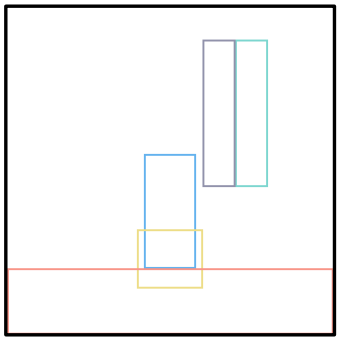
A barefoot man in striped shorts A playful black and white dog A spinning white frisbee midair		

A beach scene captures a man, clad in blue and white striped swim shorts, standing barefoot on the warm, golden sand. To his side, a playful black and white dog, with its gaze fixed on an object in the sky, waits in anticipation. Suspended in the air above them is a spinning white frisbee, creating a dynamic moment of play and excitement just off the coast, where the gentle waves lap at the shore.

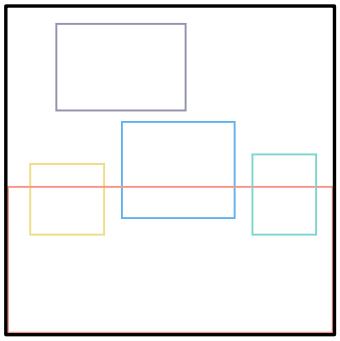
A modern flat screen TV A large decorative clock A first leather chair A second leather chair A classic fireplace A bookshelf with many DVDs		

A photo of someones living room complete with a bookshelf full of dvds, two leather chairs, a flat screen tv, fireplace, and a overly large decorative clock.

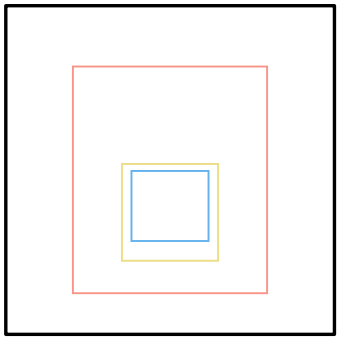
Figure 6. Qualitative results on DPG.

Instance descriptions	Layout	Image
A balancing surfer A bright yellow surfboard A cresting ocean wave A reflective glass skyscraper A beige brick skyscraper		

an individual balancing on a bright yellow surfboard, riding the crest of an ocean wave. parallel to the shore, a series of tall buildings stand in close proximity to one another, creating a dense urban skyline. the closest building has a reflective glass facade, while the one alongside it features beige brickwork.

A dull-gleaming cast-iron kettle A floral ceramic teapot A rough-hewn wooden table A woven dried-flower basket A partially-drawn curtained window		

An old-fashioned kitchen setting with a cast-iron kettle and a ceramic teapot sitting atop a rough-hewn, wooden table that bears the marks and patina of age. The kettle's metallic surface has a dull gleam, reflecting the warm ambient light, while the teapot, adorned with a floral pattern, adds a touch of nostalgia to the setting. In the background, there is a window with curtains partially drawn, allowing for a soft natural light to fill the room. Nearby, a woven basket filled with dried flowers accentuates the rustic charm of the cozy interior.

Several colorful vanilla-specked ice cream scoops A clear glass bowl A warmly lit microwave interior		

Inside the microwave sits a clear glass bowl, filled to the brim with scoops of colorful ice cream with visible flecks of vanilla beans. The microwave's interior light casts a warm glow on the ice cream, which threatens to melt if the door were to remain closed for long. It's an odd place for a cold dessert that's usually served at a chilly temperature to avoid its creamy contents from turning into a soupy mess. The microwave is positioned on a countertop, surrounded by assorted kitchen gadgets and a spice rack full of various seasonings.

Figure 7. Qualitative results on DPG.

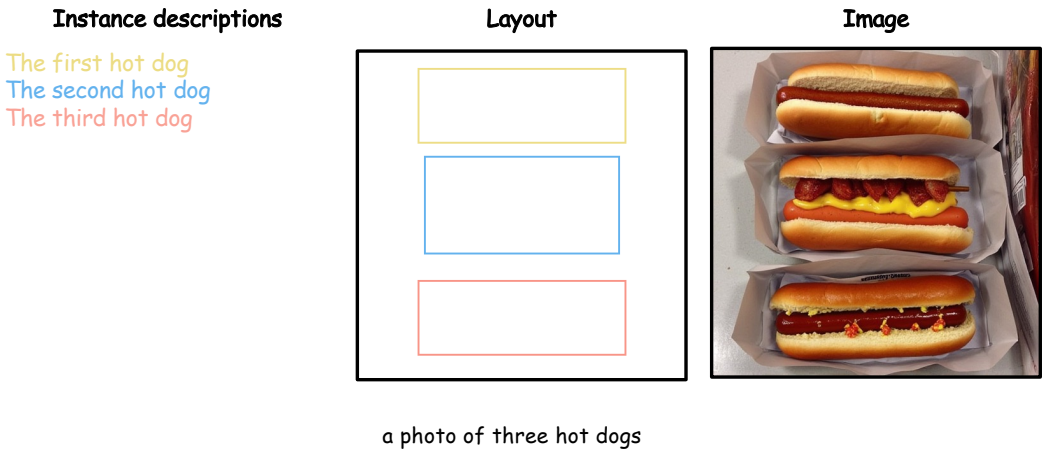


Figure 8. Qualitative results on Geneval.