Leveraging Panoptic Scene Graph for Evaluating Fine-Grained Text-to-Image Generation

Supplementary Material

A. Benchmark Construction

A.1. Dataset statistics

We follow the category counting settings in DSG [6] to report dataset statistics for our PSG-Bench. Fig. 3 shows that our proposed PSG-Bench provides a two-level category scheme, with 5 level-1 categories and 14 level-2 categories. Compared to DPG-Bench [21], PSG-Bench has a larger portion of 'Relation' due to the fact that we leverage scene graphs to generate synthetic prompts. While compared to DSG-1K [6], we has a similar portion of 'Relation'; both are around 24%, but we have a large portion of 'Action' to illustrate the relationship. Tab. 5 shows the keywords summary for our proposed dataset.

A.2. Challenge Types

We follow the challenge types design in PartiPrompts [47]. PartiPrompts is created with novel prompts and manually curating samples from recent papers. To simplify model analysis, each prompt is assigned one primary category and challenge aspect, even when multiple labels could apply. For instance, prompts containing proper nouns are categorized under WORLD KNOWLEDGE (*e.g.*, "a painting of a street in Paris") rather than ARTS. We also prioritize underrepresented categories, such as ARTS, over more common ones like ANIMALS (*e.g.*, "A raccoon wearing formal clothes, holding a cane, and a garbage bag, painted in the style of Vincent van Gogh"). The PEOPLE category always takes precedence, ensuring prompts like "a team playing baseball at the beach" are classified under PEOPLE instead of OUTDOOR SCENES, facilitating fairness and bias analysis. Challenging includes the remaining 7 challenge aspects such as IMAGINATION, QUANTITY, COMPLEX, and LINGUISTIC STRUCTURES." Tab. 6 shows the distributions of the challenge types.

A.3. Synthetic Text Prompt Template

We provide the diverse object names that have been used in our synthetic text prompts design as below:

Object Names for COCO 133 Panoptic Classes: Things (80): person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, backpack, umbrella, handbag, tie, suitcase, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, bottle, wine glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, couch, potted plant, bed, dining table, toilet, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy bear, hair drier, toothbrush.

Stuff (53): sky, grass, ground, road, sidewalk, building, wall, fence, tree, plant, water, mountain, sand, snow, floor, ceiling, rug, platform, field, earth, rock, house, bridge, tunnel, curtain, door, window, cabinet, counter, shelf, desk, table, board, pillow, blanket, clothes, sign, picture, mirror, floor-wood, floor-tile, floor-stone, floor-carpet, wall-brick, wall-stone, wall-tile, wall-wood, wall-concrete, wall-panel, wall-paper, wall-plaster, wall-plywood.

Object Names for ADE20K 150 Panoptic Classes: wall, building, sky, floor, tree, ceiling, road/route, bed, window, grass, cabinet, sidewalk/pavement, person, earth/ground, door, table, mountain/mount, plant, curtain, chair, car, water, painting/picture, sofa, shelf, house, sea, mirror, rug, field, armchair, seat, fence, desk, rock/stone, wardrobe/closet/press, lamp, tub, rail, cushion, base/pedestal/stand, box, column/pillar, signboard/sign, chest of drawers/bureau/dresser, counter, sand, sink, skyscraper, fireplace, refrigerator/icebox, grandstand/covered stand, path, stairs, runway, case/display case/showcase/vitrine, pool table/billiard table/snooker table, pillow, screen door/screen, stairway/staircase, river, bridge/span, bookcase, blind/screen, coffee table, toilet/commode/stool/throne, flower, book, hill, bench, countertop, stove, palm/palm tree, kitchen island, computer, swivel chair, boat, bar, arcade machine, hovel/hut/shack/shanty, bus, towel, light, truck, tower, chandelier, awning/sunshade/sunblind, street lamp, booth, tv, plane, dirt track, clothes, pole, land/ground/soil, bannister/balustrade/handrail, escalator/moving staircase, ottoman/pouf/hassock, bottle, buffet/counter/sideboard, poster/placard/notice/bill, stage, van, ship, fountain, conveyor belt, canopy, washer/washing machine, plaything/toy, pool, stool, barrel/cask, basket, falls, tent, bag, minibike/motorbike, cradle, oven, ball, food/solid food, step/stair, tank/storage tank, trade name, microwave, pot, animal, bicycle, lake, dishwasher, screen, blanket/cover, sculpture, hood/exhaust hood, sconce, vase, traffic light, tray, trash can, fan, pier, CRT screen, plate, monitor, bulletin board, shower, radiator, glass/drinking glass, clock, flag.

Spatial Relationships: We define spatial relationships using terms like "on the side of", "next to", "near", "on the left of", "on the right of", "on the bottom of", and "on the top

Topic	Ratio	Keywords
Position	15.12%	top, right, left, bottom, section, words, text, picture, image, icon
Content	14.79%	image, various, wall, text, several, background, includes, shows, poster, including
Signs	14.50%	sign, words, picture, shows, signs, right, left, large, image, building
Colors	13.54%	text, white, background, black, blue, image, red, letters, green, yellow
Community	8.29%	success, school, words, community, conference, services, people, information, university, program
Display	4.88%	screen, shows, displaying, image, words, digital, options, code, display, menu

Table 5. **Topic-wise Keyword Distribution with Ratios for PSG-Bench text prompts.** Our PSG-Bench primarily focus on the positioning, visual content, and design elements like color—alongside the contexts in which it is displayed.

of". The two nouns are randomly selected from three categories: persons (*e.g.*, man, woman, girl, boy), animals (*e.g.*, cat, dog, horse, rabbit, turtle), and objects (*e.g.*, table, chair, car, cup, computer). For relationships involving left, right, bottom, and top, we create contrastive prompts by swapping the nouns, such as "a girl on the left of a horse" vs. "a horse on the left of a girl".

Non-Spatial Relationships: Non-spatial relationships describe interactions between objects. We use GPT-4o to generate prompts involving verbs like "watch", "speak to", "wear", "hold", "look at", "talk to", "play with", "walk with", "stand on", and "sit on", paired with arbitrary nouns.

Attribute: For attribute, we do not limit the vocabulary to prompt GPT-4o. We ask the GPT-4o to bind the [ATTRIBUTE] to [OBJECT] in the prompt template.

Template: We first ask GPT-4o to convert the template from GenEval [13] to scene graph template. Then we can conduct the combinations and ask GPT-4o with text prompt "Could you please generate TEXT PROMPT by the given scene graph [TEMPLATE] while considering [CHALLENGE]?" Then we just need to provide the randomly picked scene graph and challenge types predefine earlier to collect synthetic prompts.

B. Implementation Details

B.1. PSG-Score Implementation

To perform Graph Matching and compute Graph Edit Distance, we use NetworkX[1] which is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. To separate the nodes from the foreground used in our proposed metric, we used saliency detection model Birefnet [50] to first extract the salient objects and later align the objects to the nodes in the scene graph.

B.2. Experimental Setup

T2I Models. The FLUX.1 [schnell] and FLUX.1 [dev] models each have 12 billion parameters and are rectified flow transformers designed for text-to-image generation. FLUX.1 [dev] offers performance similar to FLUX.1 [pro] and is available with open weights. To use their models we download their model checkpoint from huggingface for local deploy. PixArt-α, a 600 million parameter transformer-based diffusion model, competes with state-of-the-art image generators, and its checkpoint is available on GitHub. DALL·E 3, the third iteration of OpenAI's text-to-image model, is known for generating high-quality images but does not have publicly available parameters or checkpoints. Similarly, SD3.5-Large is a 8 billion parameter model and a public model checkpoint is available on huggingface.

C. Ablation Study

In this section, we ablate the key components used in our proposed evaluation metric via diverse experiments.

Object Detectors. We generate images using four T2I models on PSG-Bench. To evaluate these generated images accurately, the detection based metrics leverage the object detector to extract attribute relying on detection results—GenEval uses CLIP for color and bounding boxes for spatial relationships. To compare detectors, we randomly sample 25% of generated images, have human raters annotate bounding boxes, and report mAP against predictions. As shown in Fig. 5, our detector outperforms GenEval's, likely because PSG-Bench 's larger vocabulary makes detection harder for GenEval's Mask2Former (COCO-80 classes), whereas our detector supports more object categories, achieving higher mAP.

Attribute and Relationship Accuracy. Since attributes and relationships of generated images are directly predicted by VLMs in our approach, we verify the accuracy of these two factors. We randomly sample 25% generated images to extract attributes and relationships using GPT-4o and open-sourced Qwen2-VL-7B [41]. We use Likert scale to ask human rater to rate the correctness of extracted attributes and relationships. Fig. 6 shows the quantitative results. Both VLMs can achieve high human agreement on attributes and relationships, validating the reliability of our approach.

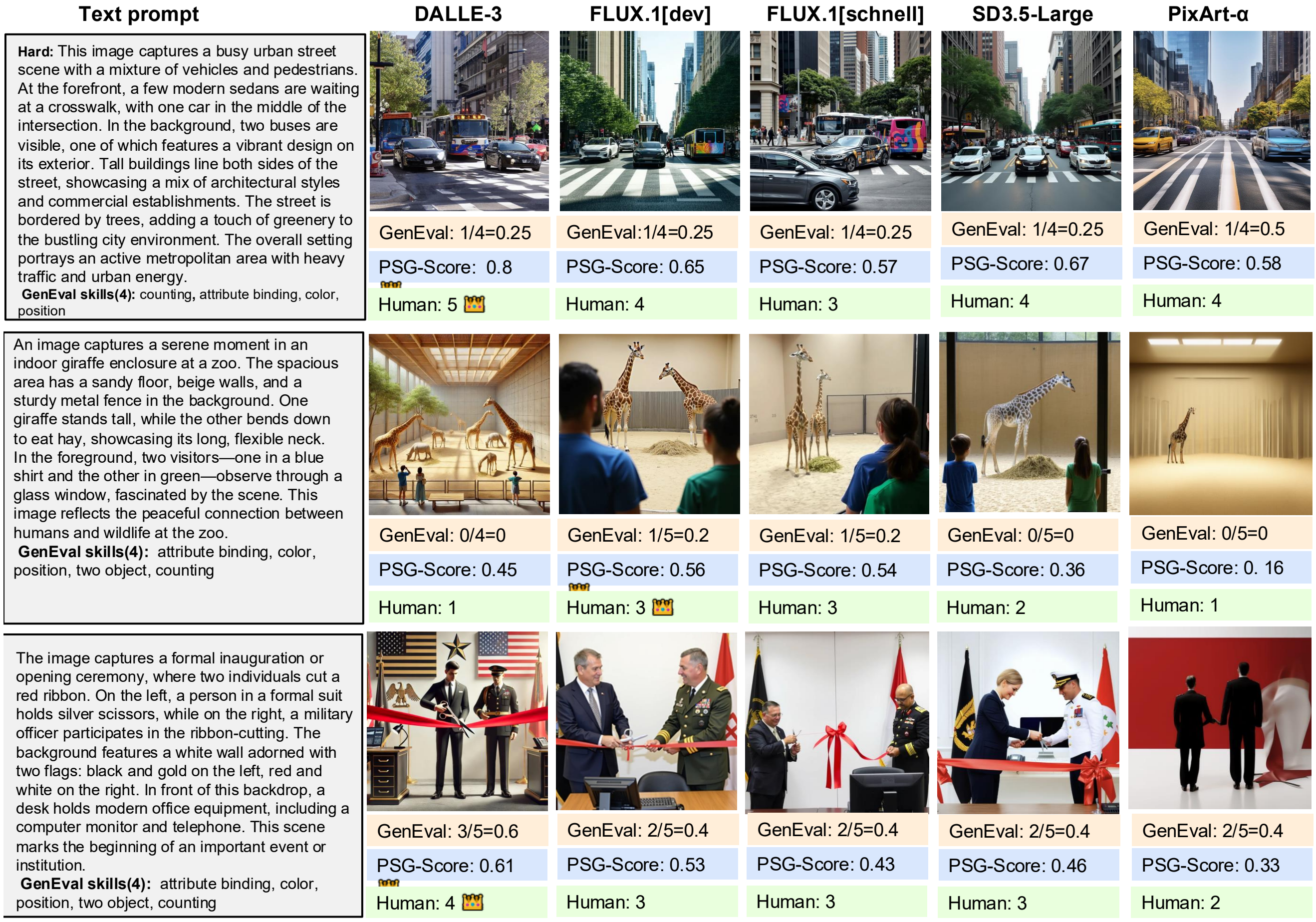
D. More Qualitative Comparison

We show more visual comparison in this section using our proposed PSG-Bench real text prompts. Fig. 9 shows the results. We compare five image generation models:DALLE-3,

[1] https://networkx.org/

Category	#Prompts	Example
IMAGINATION	525	A cat is flying.
QUANTITY	591	5 apples are placed on the dining table.
COMPLEX	776	A Renaissance-style oil painting of a deer resting beneath an ancient oak tree.
LINGUISTIC STRUCTURES	125	A t-shirt with 'NERD'
WORLD KNOWLEDGE	24	Eiffel Tower in Paris
ABSTRACT	109	'300'; '101'; 'the golden ratio'; the 'Fibonacci number'
ANIMALS	487	A woolly mammoth standing on an iceberg; Two cats gathered around a chessboard
ARTS	12	A Renaissance-style oil painting of a horse resting beneath an ancient oak tree.
ILLUSTRATIONS	32	a schematic of neural pathways;
OUTDOOR SCENES	786	Two giraffes are eating grass in a zoo.
PEOPLE	877	A formal inauguration or opening ceremony, where two individuals cut a red ribbon.
VEHICLES	656	This image captures a busy urban street scene with a mixture of vehicles and pedestrians.

Table 6. **Text Prompt Challenge Categories.** The text prompts from PSG-Bench are ranging from imaginative and complex scenarios to real-world topics like animals, vehicles, and outdoor scenes. The most common categories are People (877 prompts), Outdoor Scenes (786), and Complex (776), while Arts and World Knowledge have the fewest prompts.



Hard: This image captures a busy urban street scene with a mixture of vehicles and pedestrians. At the forefront, a few modern sedans are waiting at a crosswalk, with one car in the middle of the intersection. In the background, two buses are visible, one of which features a vibrant design on its exterior. Tall buildings line both sides of the street, showcasing a mix of architectural styles and commercial establishments. The street is bordered by trees, adding a touch of greenery to the bustling city environment. The overall setting portrays an active metropolitan area with heavy traffic and urban energy.
GenEval skills(4): counting, attribute binding, color, position

An image captures a serene moment in an indoor giraffe enclosure at a zoo. The spacious area has a sandy floor, beige walls, and a sturdy metal fence in the background. One giraffe stands tall, while the other bends down to eat hay, showcasing its long, flexible neck. In the foreground, two visitors—one in a blue shirt and the other in green—observe through a glass window, fascinated by the scene. This image reflects the peaceful connection between humans and wildlife at the zoo.
GenEval skills(4): attribute binding, color, position, two object, counting

The image captures a formal inauguration or opening ceremony, where two individuals cut a red ribbon. On the left, a person in a formal suit holds silver scissors, while on the right, a military officer participates in the ribbon-cutting. The background features a white wall adorned with two flags: black and gold on the left, red and white on the right. In front of this backdrop, a desk holds modern office equipment, including a computer monitor and telephone. This scene marks the beginning of an important event or institution.
GenEval skills(4): attribute binding, color, position, two object, counting

Figure 9. **T2I Model Ranking with Real Text Prompts.** PSG-Score correlates more closely with human preferences than GenEval, especially in complex scenes involving people and animals. Notably, DALL-E 3 consistently receives the highest PSG-Score and human ratings, while GenEval scores remain low and uninformative across most examples.

FLUX.1[dev], FLUX.1[schnell], SD3.5-Large, and PixArt-α—across three different metrics using GenEval scores, PSG-Score (ours), and human ratings. Across all rows, PSG-Score correlates more closely with human preferences than GenEval, especially in complex scenes involving people and animals. Notably, DALL·E 3 consistently receives the highest PSG-Score and human ratings, while GenEval scores remain low and uninformative across most examples. This highlights the superior alignment of PSG-Score with human judgment for evaluating prompt grounding and image quality. When the length of the text prompts increases, we can observe that PSG-Score can still align the human ratings which achieves the same conclusion for those in synthetic prompts.