

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

Xueqing Deng Linjie Yang Qihang Yu Chenglin Yang Liang-Chieh Chen
ByteDance Seed

[Project Page](#)

xueqingdeng@bytedance.com

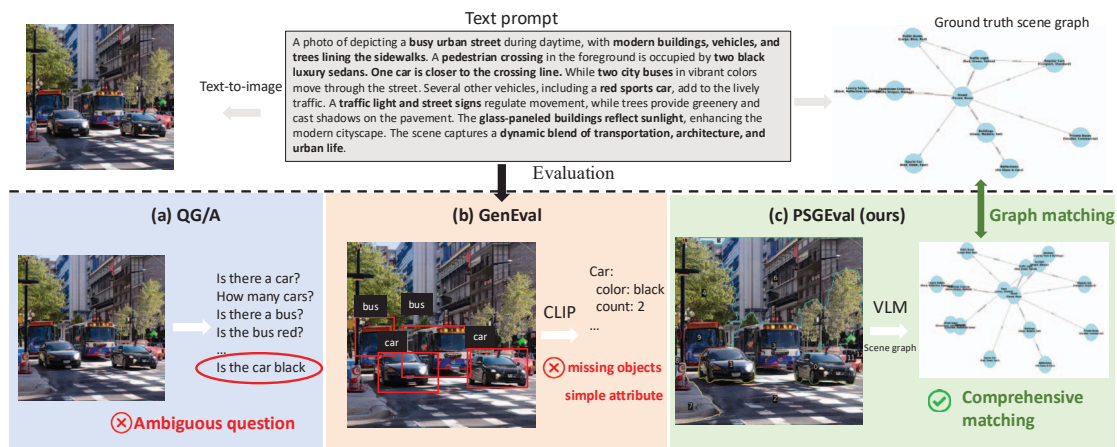


Figure 1. **Evaluation Comparison.** Evaluating complex text prompts containing multiple objects presents challenges that existing metrics fail to address adequately: (a) Question Generation/Answering (QG/A) [22] may produce ambiguous questions. For example, in the generated images, two similar cars appear, making the question “Is the car black?” confusing for vision-language models (VLMs), leading to unreliable answers. (b) GenEval [13] is restricted by the 80 object classes in COCO [32]. Objects outside this predefined set remain undetected, limiting further attribute extraction and evaluation. (c) Our proposed PSG-Score effectively detects objects, attributes, and relationships, matching them to the ground truth scene graph. This enables a more comprehensive evaluation of generated images.

Abstract

Text-to-image (T2I) models have advanced rapidly with diffusion-based breakthroughs, yet their evaluation remains challenging. Human assessments are costly, and existing automated metrics lack accurate compositional understanding. To address these limitations, we introduce PSG-Bench, a novel benchmark featuring 5K text prompts designed to evaluate the capabilities of advanced T2I models. Additionally, we propose PSG-Score, a scene graph-based evaluation metric that converts generated images into structured representations and applies graph matching techniques for accurate and scalable assessment. PSG-Score is a detection based evaluation metric without relying on QA generations. Our experimental results demonstrate that PSG-Score aligns well with human evaluations, mit-

igating biases present in existing automated metrics. We further provide a detailed ranking and analysis of recent T2I models, offering a robust framework for future research in T2I evaluation.

1. Introduction

Text-to-image (T2I) models have experienced a surge in popularity in recent years. Since the introduction of DALL-E [35], rapid advancements in diffusion models [33, 36] have led to the development of more sophisticated T2I models, such as DALL-E3 [39] and Stable Diffusion [10, 14, 33, 36]. These models have been widely adopted in creative applications like digital arts and research domains such as training data generation. Additionally, research on parameter-efficient fine-tuning techniques [20] has fueled the proliferation of new models and fine-tuned variations.

Dataset Name	#Prompts	Avg. Words	Evaluation Metric
GenEval [13]	553	8	Detection
DrawBench [37]	500	10	-
TIFA-160 [22]	160	8	QA
T2I-CompBench [23]	2400	9	Detection & QA
GenAI-Bench [27]	1600	14	QA
Gecko2K [42]	2400	9	QA
PartiPrompts [46]	1600	9	-
DSG-1K [6]	1000	17	QA
DPG-Bench [21]	1065	67	QA
EvalMuse-40K [15]	4000	67	QA
PSG-Bench (ours)	5000	113	Detection & Scene Graph

Table 1. **Dataset and Metric Comparison.** Our proposed dataset comprises 2K synthetic and 3K real-world text prompts. Additionally, we introduce PSG-Score, a novel evaluation metric grounded in object detection and scene graph analysis.

Currently, the evaluation of T2I models relies on human preference comparisons [43], which are costly and difficult to scale up. As the number of T2I models continues to grow, manual evaluation becomes impractical, necessitating the development of reliable automated evaluation methods. However, existing automated metrics struggle to assess compositional capabilities and to provide fine-grained analysis. For example, Frechet Inception Distance (FID) [17] evaluates image quality but does not consider prompt adherence, while CLIPScore [16] only produces a holistic score for an image-text pair using CLIP features [34].

Recent efforts have introduced more comprehensive metrics for T2I evaluation. One typical benchmark is GenEval [13], which assesses compositional accuracy, and semantic alignment. Specifically, it leverages CLIP [34] to extract color features and employs Mask2Former [4] as an object detector to verify object presence and alignment with the text prompt. However, GenEval has several limitations: (1) its evaluation is restricted to 80 COCO object classes [32], limiting its ability to assess diverse prompts, and (2) it only evaluates color attributes without considering more complex material attributes.

Beyond detection-based metrics, another category of evaluation methods employs Question Answering (QA) techniques [1]. These methods generate questions based on the text prompt and use Vision-Language Models (VLMs) to answer them based on the generated images [6, 22]. However, designing comprehensive questions that fully cover the prompt remains a challenge. For complex prompts, there could be ambiguous questions that are not answerable with “Yes” or “No” responses (see Fig. 1).

In addition, a common pitfall of existing metrics is that they do not penalize for extra objects generated in the image. For detection-based approach, once the entities in the prompt match with detected objects, it is considered a correct generation without accounting for extra objects. For QA-based approach, the questions generated are only related to objects in the prompt and do not care about other categories. For example, if the text prompt is about a cat

and the generation includes a cat and a dog. The QA probing will only be about the cat and no penalty will be added due to the incorrectly generated dog.

To address limitations in text prompts and evaluation methodologies, we propose: (1) PSG-Bench, a novel evaluation dataset with more challenging text prompts for advanced T2I models, and (2) PSG-Score, a panoptic scene graph-based evaluation metric combining a detection-based approach and scene-parsing capability of VLMs. PSG-Bench provides 5K text prompts, including 2K synthetic and 3K real-world prompts. It evaluates key model capabilities such as color accuracy, object counts, spatial relationships, text rendering, and unconventional object interactions. In addition, PSG-Bench includes concepts of background, for example, scene summary and background scene objects. We also incorporate human efforts to select the hard and difficult text prompts based on the model results.

To accurately measure the image-text alignment under the complex prompts, generated images are converted into scene graphs to match with ground-truth scene graphs with graph matching techniques. We provide manually annotated scene graphs for a fraction of PSG-Bench as a ground truth reference. Thanks to the graph matching approach, our metric PSG-Score is able to address the issues of ambiguous questions and proper penalties for additional objects by explicit matching of prompted and generated object entities.

We summarize our contributions as follows:

- **PSG-Bench:** A challenging evaluation **dataset** featuring 5K complex text prompts (2K synthetic + 3K real-world) for advanced T2I models.
- **PSG-Score:** A novel scene graph-based **evaluation metric** that comprehensively assesses both foreground and background content generation. We improve the detection base metric using scene graph to provide accurate assessment without using any Question Generation/Answering methods avoiding the ambiguity in question generation and vision language model bias when answering these questions.
- **Comprehensive Analysis:** We rank modern T2I models based on PSG-Score and conduct human alignment studies to validate the effectiveness of our metric.

2. Related Work

2.1. Evaluation Benchmark

Given the heavy cost and low scalability of human evaluation, automated methods are critical for evaluating the increasingly large number of new image generation models.

EvalMuse-40K [15] collects 2K real prompts from DiffusionDB [41] and 2K synthetic prompts. PartiPrompts [46] is a rich set of over 1600 prompts in English that can be used to measure model capabilities across various categories and challenge aspects. Even though it provides comprehen-

sive coverage of the evaluation aspects, it heavily relies on human evaluation. DSG-1K [6] collects prompts from multiple sources including counting/relation/text-focused prompts, to form a comprehensive evaluation set. We aim to provide a benchmark with more challenging text prompts than existing benchmarks. The challenge mainly comes from two aspects: (1) We construct synthetic text prompts using counter-factual concepts or relationships. (2) We construct long and detailed real-world text prompts harvested from detailed image captions.

2.2. Evaluation Metrics

The text-to-image community has mainly used two types of automated evaluation metrics: image quality and image-text alignment. For image quality, Inception Score [38] and Frechet Inception Distance (FID) [17] are the metrics most commonly used. They use the features of a pre-trained image classifier such as Inception v3 [40] to measure the diversity and visual fidelity of the generated images. Early approaches for measuring image-text alignment include text-image embedding similarity scores from multimodal encoders (*e.g.*, CLIPScore [16] and text similarity based on image captioning [19]). However, these metrics are all holistic scores which are not able to evaluate the fine-grained generation details of the strong T2I models. Recently, two categories of metrics have been proposed to provide comprehensive and detailed evaluation, one is based on pre-trained recognition models and the other is based on detailed VQA evaluations.

Detection-based Metrics. SOA [18] and DALL-Eval [5] employ object detection models to determine if objects, attributes, and relations in the text input are in the generated image. However, this approach only works on synthesized text inputs and measures faithfulness on limited axes (object, counting, color, and spatial relation), missing elements such as material, shape, activities, and context. The commonly used metric GenEval [13] only investigates 80 object classes [32] with 11 colors, and 4 spatial relationships. Through our experiments, we found that GenEval does not fully reflect relative performances of recent state-of-the-art models using its limited object vocabulary from its detector and simple text prompts. We therefore propose PSG-Bench which provides a more comprehensive evaluation testbed with 300 object classes including both foreground and background objects. By providing the more challenging testbed, the evaluation metric also needs to be improved due to multi-instance mixing issues (Fig. 1). Our proposed PSG-Score utilizes a novel scene-graph-based metric to resolve the ambiguity in matching generated object instances to entities in the text prompt.

QG/A-based Metrics. Question Answering (QA)-based evaluation has long been used in text summarization to assess the retention of key information via automated question

generation approaches [22]. In the multimodal domain, as large pre-trained foundation models have advanced, a growing body of research has explored verification of image-text alignment using Visual Question Answering (VQA) generated from prompts. This approach, collectively referred to as Question Generation and Answering (QG/A), enables fine-grained evaluation including verification of object categories and corresponding relations and attributes. TIFA [22] generates questions across semantic categories (*e.g.*, color, shape, counting) using GPT-3 and validates them with VQA modules such as mPLUG [29]. Yarom *et al.* take a similar approach, employing VQ²A [2] as the question generator while enhancing VQA model-human correlation through data synthesis and high-quality negative sampling. DSG-1K [6] builds on the QG/A paradigm but draws inspiration from formal semantics to address key reliability challenges identified in prior work, such as duplicated and non-atomic questions. T2I-CompBench [23] uses multiple models to evaluate different dimensions of image-text alignment: BLIP-VQA [30] for attribute binding evaluation, UniDet [50] for spatial relationship evaluation, and MiniGPT4-CoT [51] is used as a potential unified metric. Tab. 1 shows the comparisons of the existing popular evaluation benchmark and the corresponding evaluation metric. Our benchmark represents the largest existing evaluation dataset in terms of the number of text prompts. Furthermore, our novel evaluation metric leverages scene graphs to enhance traditional detection-based evaluation.

3. PSG-Bench: A Challenging T2I Benchmark

Our PSG-Bench provides both a challenging evaluation set and a new T2I evaluation metric. It features 5K challenging text prompts harvested from synthetic and real-world sources. Each prompt is annotated with a ground-truth scene graph, challenge types and a difficulty level.

Synthetic Text Prompts from Compositional Scene Graph. For synthetic text prompts, we adopt a task-specific template approach similar to GenEval. The templates include placeholders for object names, attributes, numbers, and relationships. Object names are drawn from the union set of 133 class names from COCO Panoptic [24] and 150 class names from ADE20K [49], with some classes renamed to remove ambiguity, *e.g.*, “mouse” to “computer mouse.” This choice is driven by the fact that these categories are the most common categories in real life, and are also widely used by detection models to produce relatively accurate detection results. Colors, spatial relationship and other attributes are automatically generated by GPT-4o.

Instead of using a fixed template, we utilize scene graph template generated from GenEval’s template using VLM to ask GPT-4o to synthesize the text prompt which meets the challenge types as revealed in PartiPrompts [46]. An example template is “ a/an [ATTRIBUTE A] [OBJECT

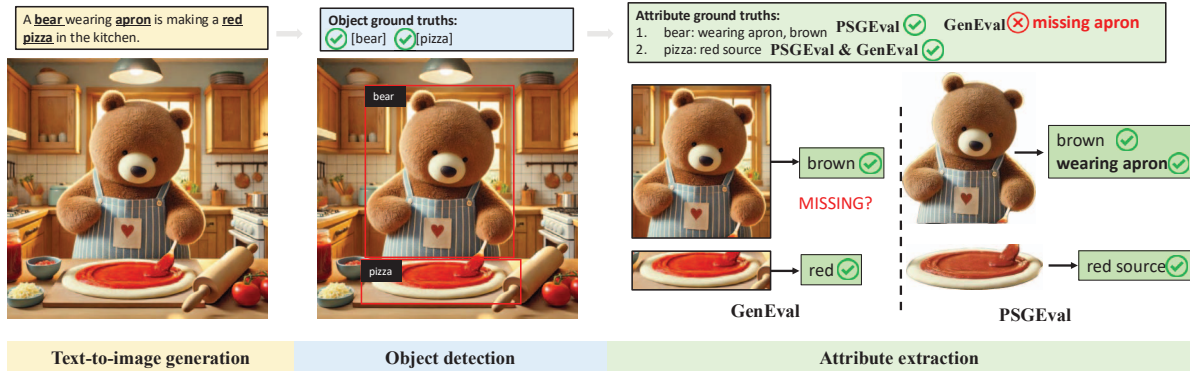


Figure 2. **Attribute Extraction Comparison.** Given a text prompt, the generated image is provided on the left. With the same object bounding boxes results from GenEval and PSG-Score shown in the middle, GenEval use CLIP to extract simple color like $\{brown, red\}$ while PSG-Score produces **accurate and comprehensive attributes** like $\{brown, apron, red\}$ thanks to the technique of Set-of-Mark and VLM (see Sec. 4.1). The significant attribute ‘apron’ is missing in GenEval attribute extraction.

A) and a/an [ATTRIBUTE B] [OBJECT B] [RELATIONSHIP] [BACKGROUND].” Fig. 7 shows the examples.

Real-world and Long Text Prompts from COCO Images. For the real-world text prompts, we ask our human raters to select 3K annotated captions from the COCONut-Pancap validation dataset [7, 8] to ensure the challenging levels of the prompts. The selection pipeline is as follows: (1) Provide the generated images from four modern image generation models, and randomly pick two images for each text prompt; (2) With the given text prompt and images, we ask human raters to select the text prompts with the images that are not perfectly aligned. In other words, if both the generated images show correct content following the text prompt, this text prompt is deemed too simple and is removed. The goal of filtering simple text prompts is to increase the performance difference of more and more powerful generative models, and to identify remaining challenges under complex prompts. Finally, all the chosen text prompts will be sent to conduct the difficulty levels categorization.

Difficulty Levels and Challenge Types. Along with the generated images and text prompts, we ask human raters to categorize the challenge types and difficulty levels. We adopt the same 11 challenge types defined in PartiPrompts [46] and 3 difficulty levels: easy, medium, and hard. Two out of four generated images from four modern T2I models are randomly picked for the evaluation of each prompt. The following question is used to ask human raters to annotate the challenge level: “Given the text prompt T and the generated images, could you please check the if the generated contents align with the text prompt? There are several suggested steps: 1) Check if the nouns (objects), actions, relationships, attributes or background (if provided) from the text prompt are generated; 2) Based on the missing content, choose the challenge level below: easy (90% contents can be generated for all images); medium (75% contents can be generated; hard (less than 50% contents can be generated).” More details of the challenge types can be

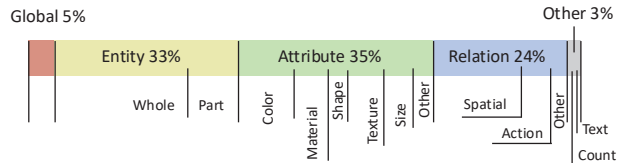


Figure 3. **Level-1 and level-2 Categories for the Keywords in Text Prompts.**

seen in Supplementary.

Scene Graph Annotation. In order to perform the PSG-Score metric, which leverages scene graphs to perform evaluation of image-text alignment, we adopt a semi-automatic annotation process. We first generate the scene graph given the text prompt using a state-of-the-art LLM (GPT-4o). Then, we provide the original text prompt and generated scene graph to human annotators, asking them to verify and correct the scene graphs to guarantee the correctness of the scene graph ground truths.

4. PSG-Score: Panoptic Scene Graph for Comprehensive T2I Evaluation

In this section, we introduce a novel evaluation metric, namely PSG-Score, which leverages panoptic scene graphs [44] as an abstraction of visual information to evaluate image-text alignment. We first introduce the generation pipeline of panoptic scene graph using panoptic segmentation models and VLMs on the generated images in Sec. 4.1. It is noted that the ground truth scene graphs of text prompts have been annotated in our PSG-Bench. Then we illustrate how to match the predicted scene graphs with the ground truth and the final score calculation in Sec. 4.2.

4.1. Scene Graph Generation

We proposed to extract a scene graph as an abstraction of the generated image for T2I alignment evaluation. We first generate panoptic segmentation masks using a state-of-the-art segmentation model [47, 48], and then use the Set-of-

Mark [45] techniques to extract scene graphs from VLMs.

Object and Scene Detection. To conduct a comprehensive scene understanding for fine-grained image-text evaluation, we propose to use panoptic segmentation models that can detect both object instances and scene segments (sky, street, *etc.*). Due to the different object categories involved in the synthetic and real-world subsets, we used two types of panoptic segmentation models for the synthetic and real-world text prompts respectively. We employ the open-vocabulary FC-CLIP [48] for synthetic set due to its rich category list, and employ the closed-vocabulary model kMaX-DeepLab [47] for real-world subset due to its better performance with COCO classes.

Extract Attributes and Relationships with VLMs. We use GPT-4o to extract the scene graphs using the image overlaid with scene masks. An example of the overlaid image is shown in Fig. 1. We design a specific prompt to enable GPT-4o to generate a scene graph with json format. The prompt is formulated as: *Could you please generate the scene graph for this caption “[caption]” by considering the overlaid mask image? The output should be formatted as follow: ’objects’:..., attributes:name:xx, value:xx, name:xx, value:xx, ’relationship’:xx:relationship:[xx].* We found that GPT-4o is able to generate correct json format in 95% of the cases. For cases that it generates incorrect grammar, we simply remove the corresponding cases from evaluation. As shown in Fig. 2, using Set-of-Mark (segmentation masks) input and VLM, we can extract more accurate attributes compared to CLIP used in GenEval [13].

Existing work [31] also adopts VLMs to generate scene graphs. They mainly focus on leveraging image-level vision understanding to extract rough scene graph for generating QA pairs for VLM model to evaluate, while we focus on accurate extraction of objects, attributes and relationships.

4.2. Evaluation Metric

The traditional evaluation metrics for scene graph generation [52] rely on the detection or segmentation results from the image. It first conducts a graph matching between the ground truth and predicted scene graph by matching object nodes. Intersection-over-Union (IoU) is often used to decide if a predicted object node is matched with a ground truth one. The same method is not applicable to image generation evaluations, as there is no ground truth object locations from the text prompt. We develop a method to match the ground truth and predicted scene graph using semantic information. In this section, we first introduce the definition and formulation of the input graphs, then introduce the similarity definition between nodes and edges, and performing graph matching based on the similarity, and finally we introduce calculation of the final metric.

Definition. By leveraging the scene graph generation method proposed in Sec. 4.1, we obtain the predicted scene

graph G_{pred} from a generated image. Besides, the ground truth scene graphs G_{gt} have been annotated on our PSG-Bench evaluation set. A scene graph G is represented with a set of nodes $V = \{v_1, v_2, \dots, v_n\}$ and their edges $E = \{e_1, e_2, \dots, e_k\}$. Each node v_i denotes an object instance detected by the segmentation model which are associated with attributes $A_i = \{a_1, a_2, \dots, a_m\}$ generated by GPT-4o. Each edge $e_i = (u_i, v_i, d_i)$ denotes a relationship d_i between two nodes u_i and v_i . The edges E are built on top of these nodes which provide the relationships between the nodes. Given a predicted scene graph G_{pred} and its ground truth scene graph G_{gt} , we conduct the graph matching algorithm first and then compute the accuracy and recall rate for the matched graphs.

Algorithm 1 Semantic Graph Matching

```

1: function EDGESIMILARITY( $w_1, w_2$ )
2:   if  $w_1 = w_2$  then return 1.0
3:   else return  $1 - \text{cosine\_distance}(g(w_1), g(w_2))$ 
4:   end if
5: end function
6:
7: function NODEMATCHING( $w_1, w_2$ )
8:   if  $\text{cosine\_distance}(g(w_1), g(w_2)) < 0.5$  then re-
9:   turn 1.0
10:  else return 0
11:  end if
12: end function
13: Input:
14: Load pre-trained word embedding model  $g$ 
15: Initialize directed graphs  $G_{gt}$  (Ground Truth) and
16:  $G_{pred}$  (Predicted)
17: Initialize  $E_{matched} = \text{Empty}(\text{Queue})$ 
18: Initialize  $V_{matched} = \text{Empty}(\text{Queue})$ 
19: Initialize  $\text{min\_edit\_dist} = 0$ 
20:
21:    $E_{matched}, V_{matched}, \text{min\_edit\_dist} =$ 
22:    $\text{GraphEditDistance}$  [26]( $G_{gt}, G_{pred},$ 
23:    $\text{NodeMatching}, \text{EdgeSimilarity}$ )

```

Node and Edge Similarity. Given the graph nodes V_{pred} , each node v_i is denoted with its object class and the corresponding attributes A_i . It is noted that each object could have multiple attributes. For example, the node for the generated sheep could be represented as $\{entity: \text{sheep}, attributes: \{\text{fluffy}, \text{happy}\}\}$. Due to the fact that a single attribute may exist in the case of different wording in the ground truth and predicted scene graphs, we propose to ex-

tract the BERT [9] embedding and then compute the cosine similarity to perform the matching with a threshold of 0.5. By perform the matching, we can obtained the matched nodes $V_{matched}$ and $V_{not_matched}$. We ablate the threshold with a small-scale experiment with 100 paired words randomly sampled from the attribute vocabulary in our dataset. We found 0.5 as a threshold is able to distinguish whether the pair of words have similar semantic meaning. Similar problem exists while trying to match the edges denoting relationships. For example, ‘on’ is equivalent to ‘above’ in most cases. Thus we adopt the same method to compute the similarity of a pair of edges between the ground truth and the predicted scene graphs.

Semantic Graph Matching. Once the node and edge similarities are obtained, we use an graph matching algorithm similar to graph edit distance [12] to find the most likely matching between the ground truth and the predicted scene graph. As shown in Algorithm 1, for each edge $e_i = (u_i, v_i, d_i), e_i \in E_{gt}$ in ground truth graph G_{gt} , we will iterate all edges from predicted graph G_{pred} : $e_j = (u_j, v_j, d_j), e_j \in E_{pred}$ and try to match $e_i \in E_{pred}$ using semantic similarity. Once the matching is done, there will be two categories of edges from E_{gt} and E_{gt} : matched edges $E_{matched}$ and not matched edges $E_{not_matched}$. In order to find the best matching, we follow the Graph Edit Distance [26] to avoid greedy matching. Graph Edit Distance (GED) is a measure of similarity between two graphs, defined as the minimum cost of transforming one graph into another through a series of edit operations such as node/edge insertion, deletion, or substitution. It provides a flexible and intuitive way to compare graph structures. We provide implementation details in our supplementary.

Final Score based on Precision and Recall. To compute the final score, we need to count the True Positive (TP), False Positive (FP) and False Negative (FN) rate from nodes $V_{matched}$ and $V_{not_matched}$ and edges $E_{matched}$ and $E_{not_matched}$. It is obvious that $TP = |E_{matched}| + |V_{matched}|$ which denotes the number of matched edges and nodes from G_{gt} and G_{pred} . FN is simple to count where there are missing predicate that exist in ground truth graph. FP is complicated, as there are often two cases involved: 1) Nodes are paired but the edge does not exist (*i.e.*, wrong edge prediction); 2) Extra nodes exist: In this case, we introduce a saliency object detector to separate the nodes in foreground and background. If the extra nodes exist in foreground, they will be counted as FP. If the extra nodes exist in background then these nodes will be ignored which means there is no penalty on these nodes since it is tolerable to have additional objects in the background that is not mentioned in the text prompt. Note that existing evaluation metrics do not penalize additional objects in the foreground using detection or QA-based approaches, which is an issue that has large impact on generation fidelity but has been largely

overlooked in the past. We summarize the rule of counting TP, FP, and FN as follows:

$$\begin{aligned}
 TP &= |E_{matched}| + |V_{matched}|, \\
 FN &= |E \subset E_{gt}, E \notin E_{pred}, E \in E_{not_matched}| \\
 &\quad + |V \subset V_{gt}, V \notin V_{pred}, V \in V_{not_matched}|, \\
 FP &= |E \subset E_{pred}, E \notin E_{gt}, E \in E_{not_matched}| \\
 &\quad + |V \subset V_{pred}, V \notin V_{gt}, V \in V_{not_matched}|, \\
 &\quad V \in V_{foreground}.
 \end{aligned} \tag{1}$$

At last, we can compute the final score which is originally F1 score with Precision and Recall as below:

$$\begin{aligned}
 Precision &= TP / (TP + FP) \\
 Recall &= TP / (TP + FN) \\
 PSG - Score_{Overall} &= 2 \times \frac{Precision \times Recall}{Precision + Recall}
 \end{aligned} \tag{2}$$

5. Experiments and Discussion

5.1. Ranking of Modern T2I Models

Models. We select 5 recently proposed T2I models to conduct the evaluation: DALL-E3 [35], FLUX.1[dev] [25], FLUX.1[schnell] [25], SD3.5-Large [11] and PixArt- α [3] to generate images on different benchmarks to conduct the model evaluation. We also conduct human evaluation using Likert scale on image-text alignment as a reference result. Tab. 2 shows the quantitative results. Fig. 7 shows the detail scores for some selected synthetic prompts.

Likert Scale on Text-to-Image Faithfulness. Annotators are asked to answer on a scale of 1 (worst) to 5 (best) to the question “Does the image match the text? Please consider a comprehensive evaluation on the nouns, attributes and relationships (spatial or action).” For each text prompt, we will provide the generated images from the aforementioned T2I models. The model names will not be provided when asking human to rate the images. Three human raters are involved in the rating then the model rating scores will be averaged across 3 people for 5K prompts.

Observation-1: GenEval can not properly reflect the model abilities using our challenging text prompts.

GenEval is based on object detection and color matching using CLIP, relying on predefined categories (*e.g.*, COCO classes). GenEval ranks SD3.5-Large higher than DALL-E3 while this does not align with human alignment likert score where DALL-E3 performs the best. As shown in the visual results from Fig. 7, DALL-E3 can generate images that provide better image-text alignment on “action”. For example, the 2nd row shows the results of “a bird is *fixing* a lamp”. According to PSG-Score, “fixing” is extracted in the relationship between the bird and the lamp which captures the critical action to reflect the true model capacity.

Method	GenEval [13]	TIFA [22]	DSG [6]	PSG-Score-Prec.	PSG-Score-Recall	PSG-Score-Overall	Likert (1~5)
FLUX.1[dev]	0.49	85.7	83.9	0.64	0.60	0.62	3.6
FLUX.1[schnell]	0.48	84.2	82.5	0.53	0.48	0.50	2.3
PixArt- α	0.46	82.1	80.2	0.62	0.53	0.57	3.3
SD3.5-Large	0.50	83.9	84.3	0.65	0.55	0.60	3.4
DALL-E3	0.48	84.9	80.1	0.70	0.67	0.64	3.9

Table 2. **Evaluation Results on PSG-Bench.** PSG-Score is more aligned to human rating scores compared to the other metrics. For different models, we report the metric evaluation on our proposed PSG-Bench. The likert score is conducted by human raters.

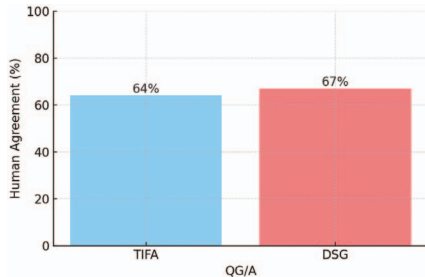


Figure 4. **Human Agreement on the yes/no QA Pairs from TIFA and DSG.** The results show low agreement with human judgments, only slightly above random guessing (50%).

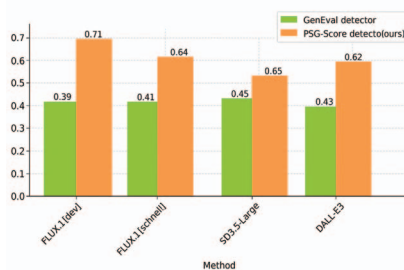


Figure 5. **Detector Comparison with mAP.** The PSG-Score detector outperforms the GenEval detector by achieving higher mAP across images generated by all models.

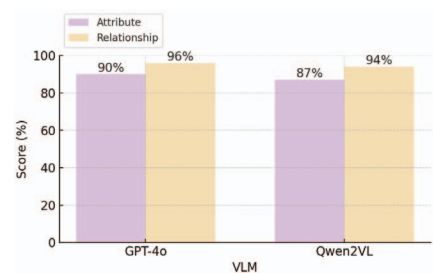


Figure 6. **Human Agreement on Attribute and Relationship.** Both GPT-4o and Qwen2VL can extract more accurate attributes with human agreement over 85%.

Method	GenEval	PSG-Score (ours)
FLUX.1[dev]	0.64	0.64
FLUX.1[schnell]	0.57	0.62
SD3.5-Large	0.62	0.65
DALL-E3	0.67	0.68

Table 3. **Comparison of GenEval and PSG-Score Metrics across Different Models using GenEval Text Prompts.** The GenEval text prompts lack sufficient complexity to effectively differentiate model performance.

Difficulty Level	Easy	Medium	Hard
GenEval	0.41/0.37	0.28/0.23	0.16/0.21
TIFA	0.52/0.39	0.33/0.29	0.24/0.17
PSG-Score (ours)	0.64/0.42	0.57/0.37	0.54/0.31

Table 4. **Correlation between Difficulty Level and Metrics Reported on PSG-Bench.** Spearman’s ρ and Kendall’s τ are reported. Our proposed metric shows stronger alignment with human ratings across three different levels of text prompts.

5.2. Metric Comparison

Implementation Details of GenEval on PSG-Bench.

GenEval is an object-focused framework to evaluate compositional image properties such as object co-occurrence, position, count, and color. It evaluates each skill as a binary classification task. For instance, a single-object evaluation prompt follows the format “a photo of a/an [OBJECT]”. If the generated image contains the object, it is labeled as correct. The dataset includes a fixed number of samples across six skill categories, where each prompt corresponds to a binary classification problem. The final skill score is calculated as $\#correct / \#total$. Since PSG-Bench prompts involve multiple skills simultaneously (*e.g.*, multiple objects, attribute binding, and relationships), they do not directly follow GenEval’s template. To compute a GenEval for PSG-Bench, we use GPT-4o to label each prompt with the GenEval skill categories. Finally, we evaluate whether the model correctly classifies all labeled skills based on GenEval’s methodology.

Observation-2: GenEval scoring FLUX.1[dev] higher score than FLUX.1[schnell] which is distilled from [dev]. FLUX.1[dev] is designed to deliver higher image quality, making it well-suited for tasks that prioritize detail and visual fidelity. In contrast, FLUX.1[schnell] is a distilled vari-

ant optimized for faster image generation. Interestingly, despite these intentions, our GenEval results as shown in Fig. 7—using challenging text prompts—revealed the opposite in performance outcomes.

Visual Comparison. As shown in Fig. 7, PSG-Score ranks DALL-E3 first, recognizing a more semantically coherent rendering of “iced” and “library” elements. For the “cat flying over clouds and mist” prompt, GenEval favors SD3.5-Large, due to better object recognition of the cat. PSG-Score again ranks DALL-E3 first, likely due to better atmospheric rendering of “clouds” and “mist.”

QG/A v.s. Detection. As shown in Tab. 2, we report two QA metrics on PSG-Bench: TIFA [22] and DSG [6]. We use GPT-4o to answer the questions for both metrics. Their metrics do not align with the human preference as well as PSG-Score on our proposed benchmark. As revealed in [28], VLMs yield bias in yes/no questions even for the SoTA GPT-4o. According to our experimental results, both TIFA and DSG generate at least one yes/no question for each text prompt. We randomly sample 1000 yes/no generated questions from TIFA and DSG, and ask human raters to rate the QA results by asking “[Generated Question] Is the given Q/A pair correct according to the image?” Fig. 4 shows the quantitative results, with both method achieving




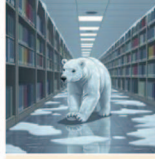



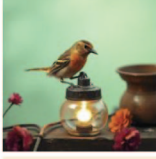







Text prompt	DALLE-3	FLUX.1[dev]	FLUX.1[schnell]	SD3.5-Large	PixArt- α
Easy: A iced bear is walking in the library . GenEval skills(4): two objects, attribute binding, color, position					
	GenEval: 1/4=0.25	GenEval: 2/4=0.5	GenEval: 1/4=0.25	GenEval: 2/4=0.5	GenEval: 2/4=0.5
	PSG-Score: 1.0 🏆	PSG-Score: 0.8	PSG-Score: 0.7	PSG-Score: 0.75	PSG-Score: 0.8
	Human: 5 🏆	Human: 4	Human: 3	Human: 4	Human: 4
Medium: A bird is fixing a lamp . GenEval skills(1): two objects					
	GenEval: 1/1=1.0	GenEval: 1/1=1.0	GenEval: 0/1=0.0	GenEval: 0/1=0.0	GenEval: 1/1=1.0
	PSG-Score: 1.0 🏆	PSG-Score: 0.66	PSG-Score: 0.34	PSG-Score: 0.56	PSG-Score: 0.66
	Human: 5 🏆	Human: 4	Human: 2	Human: 3	Human: 4
Hard: A sheep is singing while holding a microphone on the stage . Another gray sheep is dancing next to it. GenEval skills(4): two objects Color, Attribute binding, position					
	GenEval: 2/4=0.5	GenEval: 3/4=0.75	GenEval: 2/4=0.5	GenEval: 2/4=0.5	GenEval: 2/4=0.25
	PSG-Score: 0.71 🏆	PSG-Score: 0.63	PSG-Score: 0.63	PSG-Score: 0.56	PSG-Score: 0.35
	Human: 4 🏆	Human: 3	Human: 3	Human: 3	Human: 2

Figure 7. **T2I Model Ranking with Synthetic Text Prompts.** GenEval and PSG-Score scores are reported to rank the generated images from different models. We found that our proposed PSG-Score aligns more closely with human preferences.

low human agreement that are less than 70%. Since yes/no questions with random guesses should achieve 50% correctness, we consider the QA results highly unreliable with the existing QG/A approaches. Fig. 8 shows an example of TIFA generated QA pairs.

Observation-3: VLM bias leads to QG/A evaluation bias, particularly in yes/no questions. As shown in the example from Fig. 8, VLM model tends to answer ‘yes’ regardless of the actual content in the generated image. We conduct experiment of human agreement on the yes/no pairs generated from TIFA [22] and DSG [6]. We randomly sample 25% of the generated images and turn the corresponding text prompts into QA questions using TIFA and DSG. We keep the questions those with yes/no options to verify the bias of the QG/A method. Fig. 4 shows the results. The agreement only achieves 64% and 67% for TIFA and DSG yes/no QA pairs which reveals significant bias in the QG/A

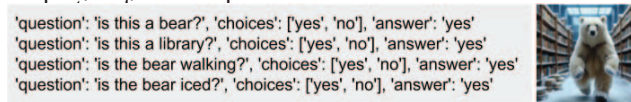


Figure 8. **Example yes/no QA from TIFA for “an iced bear is walking in the library.”** The VLM tends to favor *yes* in its responses, leading it to incorrectly answer *yes* to the question “Is the bear iced?”—even though the correct answer, based on the generated image, should be *no*.

Human Agreement on Difficulty Comparison. We com-

pare Spearman’s ρ and Kendall’s τ correlation between metrics and human Likert scores on 25% randomly sampled data from PSG-Bench. We select GenEval, TIFA to compare with our PSG-Score. As shown in Tab. 4, as difficulty increases, neither GenEval nor TIFA provides convincing evaluation as correlation with human ratings decreases.

Detection Based Metrics on Simple Text Prompts.

We conduct experiments on evaluating the models using GenEval and our proposed PSG-Score on GenEval dataset. GenEval dataset includes 533 simple text prompts with an average of 10 words. Tab. 3 shows the results. Both metric scores show limited variation across models (ranging from 0.57 to 0.68). This indicates the GenEval text prompts lack sufficient complexity to effectively differentiate model performance, as the scores remain similar across both evaluation metrics.

6. Conclusion

We propose an challenging evaluation benchmark namely PSG-Bench to evaluate the SoTA T2I models. In order to provide comprehensive and accurate evaluation, we propose PSGEval which formulates the T2I alignment metric into a graph matching problem, enabling more accurate assessment of long prompts, complex compositional structures, and extra entities. We believe the benchmark will serve as a powerful testbed for future image generation models.

References

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 2
- [2] Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. In *NAACL*, 2022. 3
- [3] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 6
- [4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2
- [5] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023. 3
- [6] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. In *ICLR*, 2024. 2, 3, 7, 8
- [7] Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. Coconut: Modernizing coco segmentation. In *CVPR*, 2024. 4
- [8] Xueqing Deng, Qihang Yu, Ali Athar, Chenglin Yang, Linjie Yang, Xiaojie Jin, Xiaohui Shen, and Liang-Chieh Chen. Coconut-pancap: Joint panoptic segmentation and grounded captions for fine-grained understanding and generation. *arXiv preprint arXiv:2502.02589*, 2025. 4
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019. 6
- [10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yan-nik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 1
- [11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 6
- [12] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and Applications*, 13:113–129, 2010. 6
- [13] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 36, 2024. 1, 2, 3, 5, 7
- [14] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. *arXiv preprint arXiv:2412.17671*, 2024. 1
- [15] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, et al. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation. *arXiv preprint arXiv:2412.18150*, 2024. 2
- [16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2, 3
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 2, 3
- [18] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE TPAMI*, 44(3):1552–1565, 2020. 3
- [19] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *CVPR*, 2018. 3
- [20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 1
- [21] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 2
- [22] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023. 1, 2, 3, 7, 8
- [23] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS*, 2023. 2, 3
- [24] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 3
- [25] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 6
- [26] Julien Lerouge, Zeina Abu-Aisheh, Romain Raveaux, Pierre Héroux, and Sébastien Adam. Graph edit distance: A new binary linear programming formulation. *arXiv preprint arXiv:1505.05740*, 2015. 5, 6
- [27] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 2
- [28] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *NeurIPS*, 2025. 7

- [29] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *EMNLP*, 2022. 3
- [30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3
- [31] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *CVPR*, 2024. 5
- [32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 3
- [33] Daniel Podell, Robin Rombach, Patrick Esser, Andreas Blattmann, and Björn Ommer. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ICLR*, 2025. 1
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2
- [35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1, 6
- [36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1
- [37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 2
- [38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 2016. 3
- [39] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. In *ACL*, 2020. 1
- [40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3
- [41] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *ACL*, 2023. 2
- [42] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. In *ICLR*, 2025. 2
- [43] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2
- [44] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *ECCV*, 2022. 4
- [45] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 5
- [46] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *TMLR*, 2022. 2, 3, 4
- [47] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means Mask Transformer. In *ECCV*, 2022. 4, 5
- [48] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *NeurIPS*, 2023. 4, 5
- [49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 3
- [50] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022. 3
- [51] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3
- [52] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, and Mohammed Benamoun. Scene graph generation: A comprehensive survey. *arXiv preprint arXiv:2201.00443*, 2022. 5