# T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-image Generation

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, Xihui Liu

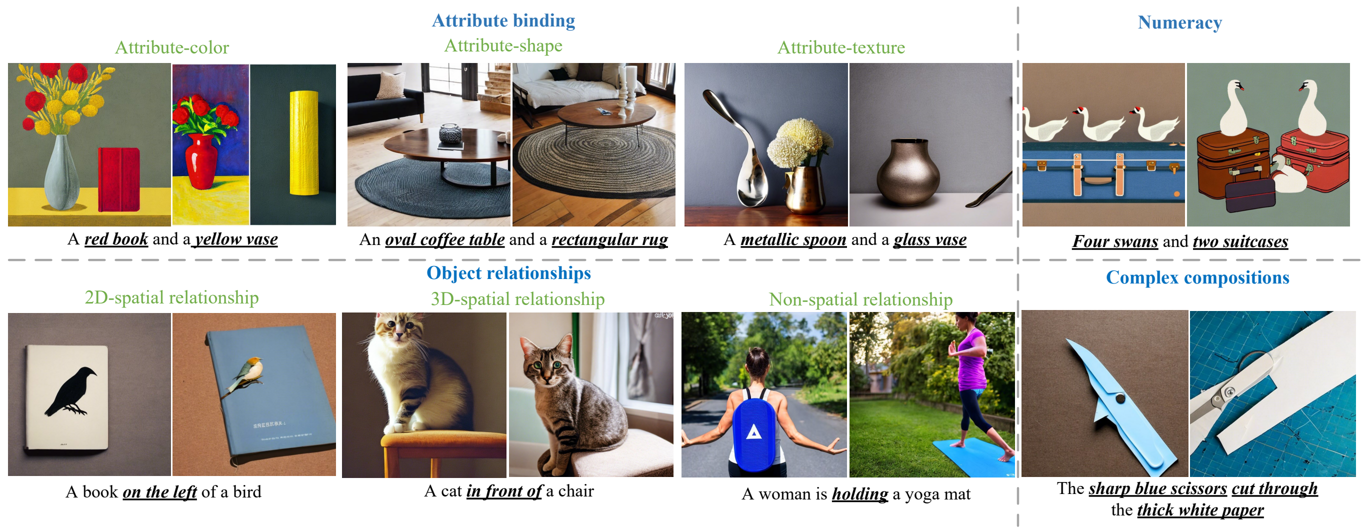


Fig. 1: **Failure cases of Stable Diffusion v2 [1]. Our compositional text-to-image generation benchmark consists of three categories: attribute binding (including color, shape, and texture), generative numeracy, object relationships (including 2D/3D-spatial relationship and non-spatial relationship), and complex compositions.**

***Abstract*—Despite the stunning ability to generate high-quality images by recent text-to-image models, current approaches often struggle to effectively compose objects with different attributes and relationships into a complex and coherent scene. We propose T2I-CompBench++, an enhanced and comprehensive benchmark for compositional text-to-image generation, consisting of 8,000 compositional text prompts from 4 categories (attribute binding, object relationships, generative numeracy, and complex compositions) and 8 sub-categories (color binding, shape binding, texture binding, 2D/3D-spatial relationships, non-spatial relationships, numeracy, and complex compositions). We further propose several evaluation metrics specifically designed to evaluate compositional text-to-image generation and explore the potential and limitations of multimodal LLMs for evaluation. We introduce a new approach, Generative mOdel finetuning with Reward-driven Sample selection (GORS), to boost the compositional text-to-image generation abilities of pretrained text-to-image models. Extensive experiments and evaluations are conducted to benchmark previous methods on T2I-CompBench++, and to validate the effectiveness of our proposed evaluation metrics and GORS approach. Project page is available at https://karine-h.github.io/T2I-CompBench-new/.**

***Index Terms*—Image generation, compositional text-to-image generation, benchmark and evaluation.**

## I. INTRODUCTION

RECENT progress in text-to-image generation [1]–[6] has showcased remarkable capabilities in creating diverse and high-fidelity images based on natural language prompts. However, we observe that even state-of-the-art text-to-image models often fail to compose multiple objects with different attributes and relationships into a complex and coherent scene, as shown in the failure cases of Stable Diffusion [1] in Figure 1. For example, given the text prompt "a blue bench on the left of a green car", the model might bind attributes to the wrong objects or generate the spatial layout incorrectly.

Previous works have explored compositional text-to-image generation from different perspectives, such as concept conjunction [7], attribute binding (focusing on color) [8], [9], and spatial relationship [10]. Most of those works focus on a sub-problem and propose their own benchmarks for evaluating their methods. However, there is no consensus on the problem definition and standard benchmark of compositional text-to-image generation. To this end, we propose a comprehensive benchmark for open-world compositional text-to-image generation, T2I-CompBench++. T2I-CompBench++ consists of

K. Huang, K. Sun and X. Liu are with The University of Hong Kong. Email: {huangky, kaiyue}@connect.hku.hk, xihuiliu@eee.hku.hk. C. Duan is with Tsinghua University. Email: duancq20@mails.tsinghua.edu.cn. E. Xie and Z. Li are with the Huawei Noah's Ark Lab. Email:{xie.enze, zhenguo}@huawei.com. Corresponding author: Xihui Liu (xihuiliu@eee.hku.hk).

four categories and eight sub-categories of compositional text prompts: (1) **Attribute binding**. Each text prompt in this category contains at least two objects and two attributes, and the model should bind the attributes with the correct objects to generate the complex scene. This category is divided into three sub-categories (color, shape, and texture) based on the attribute type. (2) **Object relationships**. The text prompts in this category each contain at least two objects with specified relationships between the objects. Based on the type of the relationships, this category consists of two sub-categories, 2D/3D-spatial relationship and non-spatial relationship. (3) **Generative numeracy**. Each prompt in this category involves one or multiple object categories with numerical quantities, ranging from one to eight. (4) **Complex compositions**, where the text prompts contain more than two objects or more than two sub-categories mentioned above. For example, a text prompt that describes three objects with their attributes and relationships.

Another challenge is the assessment of compositional text-to-image models. Most previous works evaluate the models by image-text similarity or text-text similarity (between the caption predicted from the generated images and the original text prompts) with CLIPScore [11], [12] or BLIP [13], [14]. However, both metrics do not perform well for compositionality evaluation due to the ambiguity and difficulty in compositional vision-language understanding. To address this challenge, we propose several evaluation metrics for different categories of compositional prompts. We propose *disentangled BLIP-VQA* for attribute binding evaluation to overcome the ambiguous attribute-object correspondences, *UniDet-based metric* for 2D/3D-spatial relationship and numeracy evaluation, and *MLLM (Multimodal Large Language Model)-based metric* for non-spatial relationship and complex compositions. We investigate the potential and limitations of MLLMs such as *MiniGPT-4* [15], *ShareGPT4V* [16], and *GPT-4V* [17] for compositionality evaluation.

Finally, we propose a new approach, *Generative mOdel finetuning with Reward-driven Sample selection (GORS)*, for compositional text-to-image generation. We finetune the state-of-the-art Stable Diffusion v2 [1] model with generated images that highly align with the compositional prompts, where the fine-tuning loss is weighted by the reward which is defined as the alignment score between compositional prompts and generated images. This approach is simple but effective in boosting the model's compositional abilities and can serve as a new baseline for future explorations.

In summary, our contributions are three-folded. (1) We propose a comprehensive benchmark for compositional text-to-image generation which consists of 8,000 prompts from 4 categories (attribute binding, object relationships, numeracy, and complex compositions) and 8 sub-categories (color binding, shape binding, texture binding, 2D/3D-spatial relationships, non-spatial relationships, numeracy, and complex compositions). (2) We propose evaluation metrics that are specifically designed for compositional text-to-image evaluation. Experiments validate that the proposed evaluation metrics are highly correlated with human perception. (3) We benchmark several previous text-to-image models on our proposed benchmark and evaluation metrics, and propose a simple and effective approach, GORS, for boosting the compositionality of text-to-image models.

## II. Related work

**Text-to-image generation.** Early works [19]–[24] explore different network architectures and loss functions based on generative adversarial networks (GAN) [25]. Recently, diffusion models have achieved remarkable success for text-to-image generation [1], [26]–[33]. By training on web-scale data, T2I models such as Stable Diffusion [1], [34], [35], DALL-E 3 [33], MDM [31], and Pixart-$\alpha$ [30], have shown remarkable generative power. Current state-of-the-art models such as Stable Diffusion [1] still struggle to compose multiple objects with attributes and relationships in a complex scene. Some recent works attempt to align text-to-image models with human feedback [36], [37]. RAFT [38] proposes reward-ranked fine-tuning to align text-to-image models with certain metrics. Our proposed GORS approach is a simpler finetuning approach that does not require multiple iterations of sample generation and selection.

**Compositional text-to-image generation.** Researchers have delved into various aspects of compositionality in text-to-image generation to achieve visually coherent and semantically consistent results [7], [8], [10], [39]–[42]. Previous work focused on concept conjunction and negation [7], attribute binding with colors [8], [9], [43], generative numeracy [44], and spatial relationships between objects [10], [45]. However, those work each target at a sub-problem, and evaluations are conducted in constrained scenarios. Recent compositional studies typically fall into two categories [46]: one relies on cross attention maps for compositional generation [47]–[49], while the other integrates layout as a generation condition [50]–[54]. LMD [44], RPG [55], and RealCompo [56] utilize LLMs or MLLMs to reason out layouts or decompose compositional problems. Our work is the first to introduce a comprehensive benchmark for compositional text-to-image generation.

**Benchmarks for text-to-image generation.** Early works evaluate text-to-image on CUB birds [57], Oxford flowers [58], and COCO [59] which are easy with limited diversity. As the text-to-image models become stronger, more challenging benchmarks have been introduced. DrawBench [3] consists of 200 prompts to evaluate counting, compositions, conflicting, and writing skills. DALL-EVAL [60] proposes PaintSkills to evaluate visual reasoning skills, image-text alignment, image quality, and social bias by 7,330 prompts. HE-T2I [61] proposes 900 prompts to evaluate counting, shapes, and faces for text-to-image. Several compositional text-to-image benchmarks have also been proposed. Park *et al.* [43] proposes a benchmark on CUB Birds [57] and Oxford Flowers [58] to evaluate the models' ability to generate images with object-color and object-shape compositions. ABC-6K and CC500 [8] benchmarks are proposed to evaluate attribute binding for text-to-image models, but they only focused on color attributes. HRS-Bench [18] is a general-purpose benchmark that evaluates 13 skills with 45,000 prompts. Compositionality is only

TABLE I: Comparison with previous compositional text-to-image benchmarks.

| Benchmark | Prompts number and coverage | Vocabulary diversity |
|---|---|---|
| CC-500 [8] | 500 attr bind (color) | 196 nouns, 12 colors |
| ABC-6K [8] | 6,000 attr bind (color) | 3,690 nouns, 33 colors |
| Attn-Exct [9] | 210 attr bind (color) | 24 nouns, 11 colors |
| HRS-comp [18] | 1,000 attr bind (color, size), 2,000 rel (spatial, action) | 620 nouns, 5 colors, 11 spatial, 569 actions |
| T2I-CompBench++ | 3,000 attr bind (color, shape, texture)<br>3,000 rel (2D/3D spatial, non-spatial)<br>1,000 numeracy<br>1,000 complex | 2,470 nouns, 33 colors, 32 shapes, 23 textures<br>7 2D-spatial rel, 3 3D-spatial rel, 875 non-spatial rel |

one of the 13 evaluated skills which is not extensively studied. We propose the first comprehensive benchmark for open-world compositional text-to-image generation, shown in Table I.

**Evaluation metrics for text-to-image generation.** Existing metrics for text-to-image generation can be categorized into fidelity assessment, alignment assessment, and LLM-based metrics. Traditional metrics such as Inception Score (IS) [62] and Frechet Inception Distance (FID) [63] are commonly used to evaluate the fidelity of synthesized images. To assess the image-text alignment, text-image matching by CLIP [11] and BLIP2 [14] and text-text similarity by BLIP [13] captioning and CLIP text similarity are commonly used. Some works leverage the strong reasoning abilities of large language models (LLMs) for evaluation [64]–[66]. Besides, human preferences or feedbacks are also included in text-to-image generation evaluation [67]–[74]. More fine-grained metrics are proposed in [75]. However, there was no comprehensive study on how well those evaluation metrics work for compositional text-to-image generation. We propose evaluation metrics specifically designed for our benchmark and validate that our proposed metrics align better with human perceptions.

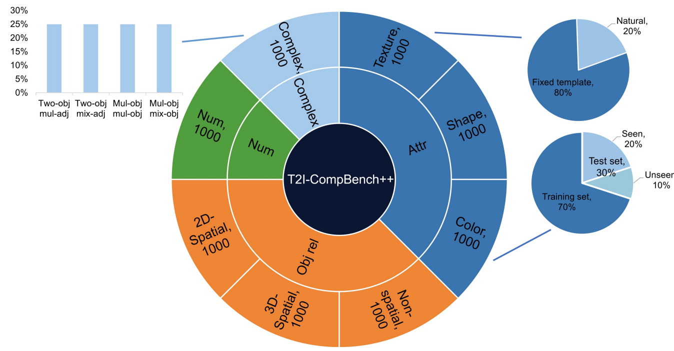


Fig. 2: Data statistics of T2I-CompBench++.

## III. T2I-CompBench++

Compositionality of text-to-image models refers to the ability of models to compose different concepts into a complex and coherent scene according to text prompts. To provide a clear definition of the problem and to build our benchmark, we introduce four categories and eight sub-categories of compositionality, attribute binding (including three sub-categories: color, shape, and texture), object relationships (including three sub-categories: 2D/3D-spatial relationship and non-spatial relationship), numeracy, and complex compositions. We generate 1,000 text prompts (700 for training and 300 for testing) for each sub-category, resulting in 8,000 compositional text prompts in total. We take the balance between seen *v.s.* unseen compositions in the test set, prompts with fixed sentence template *v.s.* natural prompts, and simple *v.s.* complex prompts into consideration when constructing the benchmark. The text prompts are generated with either predefined rules or ChatGPT [76], so it is easy to scale up. Comparisons between our benchmark and previous benchmarks are shown in Table. I, and data statistics in Figure 2.

### A. Attribute Binding

A critical challenge for compositional text-to-image generation is attribute binding, where attributes must be associated with corresponding objects in the generated images. We find that models tend to confuse the association between attributes and objects when there are more than one attribute and more than one object in the text prompt. For example, with the text prompt "A room with blue curtains and a yellow chair", the text-to-image model might generate a room with yellow curtains and a blue chair. We introduce three sub-categories, color, shape, and texture, according to the attribute type, and construct 1000 text prompts for each sub-category. For each sub-category, there are 800 prompts with the fixed sentence template "a {adj} {noun} and a {adj} {noun}" (*e.g.*, "a red flower and a yellow vase") and 200 natural prompts without predefined sentence template (*e.g.*, "a room with blue curtains and a yellow chair"). The 300-prompt test set of each sub-category consists of 200 prompts with seen adj-noun compositions (adj-noun compositions appeared in the training set) and 100 prompts with unseen adj-noun compositions (adj-noun compositions not in the training set).

**Color.** Color is the most commonly-used attribute for describing objects in images, and current text-to-image models often confuse the colors of different objects. The 1,000 text prompts related to color binding are constructed with 480 prompts from CC500 [8], 200 prompts from COCO [59], and 320 prompts generated by ChatGPT.

The prompt for ChatGPT is: *Please generate prompts in the format of "a {adj} {noun} and a {adj} {noun}" by using the color adj. , such as "a green bench and a red car".*

**Shape.** We define a set of shapes that are commonly used for describing objects in images: long, tall, short, big, small, cubic, cylindrical, pyramidal, round, circular, oval, oblong, spherical, triangular, square, rectangular, conical, pentagonal, teardrop, crescent, and diamond. We provide those shape attributes to ChatGPT and ask ChatGPT to generate prompts by composing those attributes with arbitrary objects, for example, "a rectangular clock and a long bench".

For fixed sentence template, the prompt for ChatGPT is: *Please generate prompts in the format of "a {adj} {noun} and a {adj} {noun}" by using the shape adj.: long, tall, short, big, small, cubic, cylindrical, pyramidal, round, circular, oval, oblong, spherical, triangular, square, rectangular, conical, pentagonal, teardrop, crescent, and diamond.* For natural prompts, the prompt for ChatGPT is: *Please generate objects with shape adj. in a natural format by using the shape adj.: long, tall, short, big, small, cubic, cylindrical, pyramidal, round, circular, oval, oblong, spherical, triangular, square, rectangular, conical, pentagonal, teardrop, crescent, and diamond.*

**Texture.** Textures are also commonly used to describe the appearance of objects. They can capture the visual properties of objects, such as smoothness, roughness, and granularity. We often use materials to describe the texture, such as wooden, plastic, and rubber. We define several texture attributes and the objects that can be described by each attribute. We generate 800 text prompts by randomly selecting from the possible combinations of two objects each associated with a textural attribute, *e.g.*, "A rubber ball and a plastic bottle". We also generate 200 natural text prompts by ChatGPT.

We generate 200 natural text prompts by ChatGPT with the following prompt: *Please generate objects with texture adj. in a natural format by using the texture adj.: rubber, plastic, metallic, wooden, fabric, fluffy, leather, glass.* Besides the ChatGPT-generated text prompts, we also provide the predefined texture attributes and objects that can be described by each texture, as shown in appendix Table X. We generate 800 text prompts by randomly selecting from the possible combinations of two objects each associated with a textural attribute, *e.g.*, "A rubber ball and a plastic bottle".

### B. Object Relationship

When composing objects in a complex scene, the relationship between objects is a critical factor. We introduce 1,000 text prompts for 2D/3D-spatial relationships and non-spatial relationships, respectively.

**2D-spatial relationships.** We use "on the side of", "next to", "near", "on the left of", "on the right of", "on the bottom of", and "on the top of" to define 2D-spatial relationships. The two nouns are randomly selected from persons (*e.g.*, man, woman, girl, boy, person, *etc.*), animals (*e.g.*, cat, dog, horse, rabbit, frog, turtle, giraffe, *etc.*), and objects (*e.g.*, table, chair, car, bowl, bag, cup, computer, *etc.*). For spatial relationships including left, right, bottom, and top, we construct contrastive prompts by swapping the two nouns, for example, "a girl on the left of a horse" and "a horse on the left of a girl".

**3D-spatial relationships.** To illustrate the 3D-spatial relationships, we use "in front of", "behind" and "hidden by" to define 3D-spatial relationships. Similar to 2D-spatial relationships, we use the same vocabulary to construct the pairs in a prompt, such as "a girl in front of a horse".

**Non-spatial relationships.** Non-spatial relationships usually describe the interactions between two objects. We prompt ChatGPT to generate text prompts with non-spatial relationships (*e.g.*, "watch", "speak to", "wear", "hold", "have", "look at", "talk to", "play with", "walk with", "stand on", "sit on", *etc.*) and arbitrary nouns.

The prompt for ChatGPT is: *Please generate natural prompts that contain subjects and objects by using relationship words such as wear, watch, speak, hold, have, run, look at, talk to, jump, play, walk with, stand on, and sit on.*

### C. Numeracy.

In the construction of numeracy dataset, we introduce 1,000 prompts, and adopt a tripartite structure based on the quantities of objects. The initial 30% part of the numeracy contains scenarios involving a singular object, followed by the subsequent 30% comprising two objects, and the remaining 40% encompassing multiple objects, with the quantities ranging from one to eight. We construct a list of objects, and randomly combining objects and quantities a standardized format: "number object" (*e.g.,* "one pear and three knives"). In order to incorporate diversity of expressions, each part is characterized by a combination of fixed templates and flexible prompts at a ratio of 4:1. The Fixed templates adhere to the predefined format, while flexible prompts incorporate natural language expressions encountered in our daily lives with multiple 'number object' included (*e.g.,* "three plates and three pens were on the table").

We construct a list of objects by asking ChatGPT: *Please generate a list of 150 types of objects and their plurals.* Following that, we manually filtered out repeated entries and uncommon objects. For generating flexible prompts, we ask ChatGPT *Convert to natural prompt: You are given number sentences, each containing multiple objects. For each sentence, use the objects in it to create a natural compositional sentence.*

### D. Complex Compositions

To test text-to-image generation approaches with more natural and challenging compositional prompts in the open world, we introduce 1,000 text prompts with complex compositions of concepts beyond the pre-defined patterns. Regarding the number of objects, we create text prompts with more than two objects, for example, "a room with a blue chair, a black table, and yellow curtains". In terms of the attributes associated with objects, we can use multiple attributes to describe an object (denoted as *multiple attributes*, *e.g.*, "a big, green apple and a tall, wooden table"), or leverage different types of attributes in a text prompt (denoted as *mixed attributes*, *e.g.*, the prompt "a tall tree and a red car" includes both shape and color attributes). We generate 250 text prompts with ChatGPT for each of the four scenarios: *two objects with multiple attributes, two objects with mixed attributes, more than two objects with multiple attributes, and more than two objects with mixed attributes*. Relationship words can be adopted in each scenario to describe the relationships among two or more objects. For each scenario, we split 175 prompts for the training set and 75 prompts for the test set.

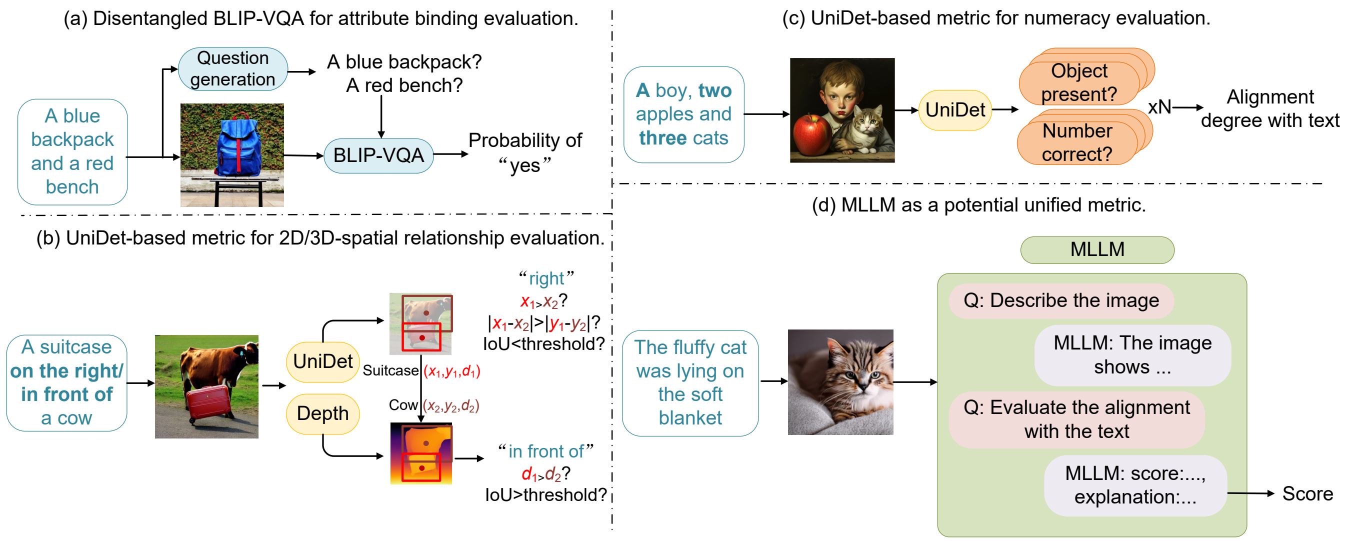


Fig. 3: Illustration of our proposed evaluation metrics: (a) Disentangled BLIP-VQA for attribute binding evaluation, (b) UniDet for 2D/3D-spatial relationship evaluation, (c) UniDet for numeracy evaluation, and (d) MLLM as a potential unified metric.

(1) For 2 objects with mixed attributes, the prompt for ChatGPT is: *Please generate natural compositional phrases, containing 2 objects with each object one adj. from {color, shape, texture} descriptions and spatial (left/right/top/bottom/next to/near/on side of) or non-spatial relationships.* (2) For 2 objects with multiple attributes, the prompt for ChatGPT is: *Please generate natural compositional phrases, containing 2 objects with several adj. from {color, shape, texture} descriptions and spatial (left/right/top/bottom/next to/near/on side of) or non-spatial relationships.* (3) For multiple objects with mixed attributes, the prompt for ChatGPT is: *Please generate natural compositional phrases, containing multiple objects (number>2) with each one adj. from {color, shape, texture} descriptions and spatial (left/right/top/bottom/next to/near/on side of) non-spatial relationships.* (4) For multiple objects with multiple attributes, the prompt for ChatGPT is: *Please generate natural compositional phrases, containing multiple objects (number>2) with several adj. from {color, shape, texture} descriptions and spatial (left/right/top/bottom/next to/near/on side of) or non-spatial relationships.*

## IV. Evaluation Metrics

Evaluating compositional text-to-image generation is challenging as it requires comprehensive and fine-grained cross-modal understanding. Existing evaluation metrics leverage vision-language models trained on large-scale data for evaluation. CLIPScore [11], [12] calculates the cosine similarity between text features and generated-image features extracted by CLIP. Text-text similarity by BLIP-CLIP [9] applies BLIP [13] to generate captions for the generated images, and then calculates the CLIP text-text cosine similarity between the generated captions and text prompts. Those evaluation metrics can measure the coarse text-image similarity, but fails to capture fine-grained text-image correspondences in attribute binding and spatial relationships. To address those limitations, we propose new evaluation metrics for compositional text-to-image generation, shown in Fig. 3. Concretely, we propose *disentangled BLIP-VQA* for attribute binding evaluation, *UniDet-based metric* for 2D/3D-spatial relationship and numeracy evaluation, and *MLLM-based metric* for non-spatial relationship and complex prompts. We further investigate the potential and limitations of multimodal large language models such as *MiniGPT-4* [15] with Chain-of-Thought [77], ShareGPT4V [16], and GPT-4V [17] for compositionality evaluation.

### A. Disentangled BLIP-VQA for Attribute Binding Evaluation

We observe that the major limitation of the BLIP-CLIP evaluation is that the BLIP captioning models do not always describe the detailed attributes of each object. For example, the BLIP captioning model might describe an image as "A room with a table, a chair, and curtains", while the text prompt for generating this image is "A room with yellow curtains and a blue chair". So explicitly comparing the text-text similarity might cause ambiguity and confusion.

Therefore, we leverage the visual question answering (VQA) ability of BLIP [13] for evaluating attribute binding. For instance, given the image generated with the text prompt "a green bench and a red car", we ask two questions separately: "a green bench?", and "a red car?". By explicitly disentangling the complex text prompt into two independent questions where each question contains only one object-attribute pair, we avoid confusion of BLIP-VQA. The BLIP-VQA model takes the generated image and several questions as input and we take the probability of answering "yes" as the score for a question. We compute the overall score by multiplying the probability of answering "yes" for each question. The proposed disentangled BLIP-VQA is applied to evaluate the attribute binding for color, shape, and texture.

### B. UniDet-based Metric for Spatial Relationships and Numeracy Evaluation

Many vision-language models exhibit limitations in reasoning spatial relationships, such as distinguishing between "left" and "right," as well as in numerical counting. Consequently, we propose the use of a detection-based evaluation metric to assess performance in spatial relationships and numeracy.

**2D-spatial relationships.** We first use UniDet [78] to detect objects in the generated image. Then we determine the spatial relationship between two objects by comparing the locations of the centers of the two bounding boxes. Denote the center of the two objects as $(x_1, y_1)$ and $(x_2, y_2)$, respectively. The first object is on the left of the second object if $x_1 < x_2, |x_1 - x_2| > |y_1 - y_2|$, and the intersection-over-union (IoU) between the two bounding boxes is below the threshold of 0.1. Other spatial relationships "right", "top", and "bottom" are evaluated similarly. We evaluate "next to", "near", and "on the side of" by comparing the distances between the centers of two objects with a threshold.

**3D-spatial relationships.** For 3D-spatial relationships, we leverage depth estimation [79], in conjunction with the detection-based metric. Let $d_1$ and $d_2$ represent the mean values of the depth maps corresponding to the two objects. Specifically, if $d_1 > d_2$ and the IoU metric exceeds the predetermined threshold of 0.5, it is inferred that the first object is positioned in front of the second object. Other 3D-spatial relationships "behind", "hidden by" are evaluated similarly.

**Numeracy.** For numeracy, we first extract the names of objects and their corresponding quantities from the prompts. Then, we leverage UniDet to detect objects in the images. This scoring mechanism is designed to consider both objects and their numerical correctness proportionally based on the number of object categories mentioned in the prompt. Let the variable $n$ represent the count of distinct object categories referenced in the given prompt. For every identified object in the image, the assigned score is calculated as $1/(2n)$ points. Furthermore, another $1/(2n)$ points are allocated if the generated quantity aligns accurately with the specified category.

### C. 3-in-1 Metric for Complex Compositions Evaluation

Since different evaluation metrics are designed for evaluating different types of compositionality, there is no single metric that works well for all categories. For non-MLLM metric, we empirically find that the Disentangled BLIP-VQA works best for attribute binding evaluation, UniDet-based metric works best for 2D/3D-spatial relationship and numeracy evaluation, and CLIPScore works best for non-spatial relationship evaluation. Thus, we design a 3-in-1 evaluation metric which computes the average score of CLIPScore, Disentangled BLIP-VQA, and UniDet, as the evaluation metric for complex compositions.

### D. MLLM-based Evaluation Metric

Multimodal Large Language Models enable users to instruct LLMs to analyze user-provided image inputs. By integrating the visual modality, MLLMs enhance the capability of language-only systems, providing them with new interfaces to address a variety of tasks. However, the multimodal abilities of MLLMs regarding the compositional text-to-image generation remain unclear. Within this context, we analyze the abilities of three types of MLLMs, namely MiniGPT-4, ShareGPT4V [16] and GPT-4V [17], focusing on their performance in compositional problems.

**Evaluation methods of MLLMs.** By aligning a pretrained visual encoder with a frozen large language model, multimodal large language models such as *MiniGPT-4* [15] have demonstrated great abilities in vision-language cross-modal understanding. But the current MiniGPT-4 model exhibit limitations such as inaccurate understanding of images and hallucination issues. *ShareGPT4V* [16] is a large-scale image-text dataset featuring 1.2 million detailed captions characterized by richness and diversity. *GPT-4V*, a state-of-the-art MLLM, is developed on the foundation of the state-of-the-art LLM, GPT-4 [80], and trained extensively on a large-scale dataset containing multimodal information [80]. Employing MLLM as an evaluation metric, we submit generated images to the model and assess their alignment with the provided text prompt. This evaluation involves soliciting predictions for the image-text alignment score.

**Prompt template for MLLM evaluation.** We leverage MLLMs as an evaluation metric by feeding the generated images to the model and asking two questions with Chain-of-Thought [77]: "describe the image" and "predict the image-text alignment score". We detail the prompts used for MLLM evaluation metric. For each sub-category, we ask two questions in sequence: "describe the image" and "predict the image-text alignment score", see appendix Table XI-XV.

## V. Boosting Compositional Text-to-image Generation with GORS

We introduce a simple but effective approach, Generative mOdel finetuning with Reward-driven Sample selection (GORS), to improve the compositional ability of pretrained text-to-image models. Our approach finetunes a pretrained text-to-image model such as Stable Diffusion [1] with generated images that highly align with the compositional prompts, where the fine-tuning loss is weighted by the reward which is defined as the alignment score between compositional prompts and generated images.

Specifically, given the text-to-image model $p_\theta$ and a set of text prompts $y_1, y_2, \cdots, y_n$, we first generate $k$ images for each text prompt, resulting in $kn$ generated images $x_1, x_2, \cdots, x_{kn}$. Text-image alignment scores $s_1, s_2, \cdots, s_{kn}$ are predicted as rewards. We select the generated images whose rewards are higher than a threshold to fine-tune the text-to-image model. The selected set of samples are denoted as $\mathcal{D}_s$. During fine-tuning, we weight the loss with the reward of each sample. Generated images that align with the compositional prompt better are assigned higher loss weights, and vice versa. The loss function for fine-tuning is

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y,s)\in\mathcal{D}_s}\Big[s \cdot \|\epsilon - \epsilon_\theta\left(z_t, t, y\right)\|_2^2\Big], \quad (1)$$

where $(x, y, s)$ is the triplet of the image, text prompt, and reward, and $z_t$ represents the latent features of $x$ at timestep $t$. We adopt LoRA [81] for efficient finetuning.

## VI. Experiments

### A. Experimental Setup

**Evaluated models.** We evaluate the performance of 6 baseline text-to-image models on T2I-CompBench++. *Stable Diffusion v1-4* and *Stable Diffusion v2* [1] are text-to-image models trained on large amount of image-text pairs. *Composable Diffusion* [7] is designed for conjunction and negation of concepts for pretrained diffusion models. *Structured Diffusion* [8] and *Attend-and-Excite* [9] are designed for attribute binding for pretrained diffusion models. We re-implement those approaches on Stable Diffusion v2 to enable fair comparisons. *GORS* is our proposed approach which finetunes Stable Diffusion v2 with selected samples and their rewards. To avoid the bias from selecting samples by evaluation metrics as reward [43], [82], we introduce new reward models which are different from our proposed evaluation metrics, denoted as *GORS-unbiased*. Specifically, we adopt Grounded-SAM [83] as the reward model for the attribute binding category. We extract the segmentation masks of attributes and their associated nouns separately with Grounded-SAM, and use the Intersection-over-Union (IoU) between the attribute masks and the noun masks together with the grounding mask confidence to represent the attribute binding performance. We apply GLIP-based [84] selection method for 2D/3D-spatial relationships and numeracy. For non-spatial relationships, we adopt BLIP [13] to generate image captions and CLIP [11], [12] to measure the text-text similarity between the generated captions and the input text prompts. For complex compositions, we integrate the 3 aforementioned reward models as the total reward. Those sample selection models are different from the models used as evaluation metrics.

We further conduct evaluations on 3 recent text-to-image models, *i.e.*, PixArt-$\alpha$ [30], Stable Diffusion XL [34] and DALL-E 3 [33]. PixArt-$\alpha$ [30] is a Transformer-based T2I diffusion model, boasting competitive image generation quality comparable to state-of-the-art image generators with low training costs. We test PixArt-$\alpha$ fine-tuned on GORS, denoted as PixArt-$\alpha$-ft. Stable Diffusion XL 1.0 [34] building upon previous iterations of Stable Diffusion models, represents a powerful text-to-image generation framework. DALLE-3 [33] is a new text-to-image generation system, enhancing the alignment between generated images and provided text description.

**Implementation details.** We employ pre-trained models for our proposed evaluation metrics. For BLIP-VQA, we utilize the BLIP w/ ViT-B and CapFilt-L [13] pretrained on image-text pairs and fine-tuned on VQA. We employ the UniDet [78] model trained on 4 large-scale detection datasets (COCO [59], Objects365 [85], OpenImages [86], and Mapillary [87]). For CLIPScore, we use the “ViT-B/32” pretrained CLIP model [11], [12]. For MiniGPT4-CoT, we utilize the Vicuna 13B of MiniGPT4 [15] variant with a temperature setting of 0.7 and a beam size of 1. For ShareGPT4V [16] and GPT-4V [17], we use the default parameters, with temperature setting of 0.2 and 1, respectively.

We implement our proposed GORS upon the codebase of diffusers [88] (Apache License), and finetune the self-attention layers of the CLIP text encoder and the attention layers of U-net using LoRA [81]. The model is trained by AdamW optimizer [89] with $\beta_1$=0.9, $\beta_2$=0.999, $\epsilon$=1$e$-8, and weight decay of 0.01. The batch size is 5. The model is trained on 8 32GB NVIDIA v100 GPUs.

### B. Evaluation Metrics

We generate 10 images for each text prompt in T2I-CompBench for automatic evaluation. To ensure a fair comparison, the images are generated using the fixed seed across all models*.

**Previous metrics.** *CLIPScore* [11], [12] (denoted as *CLIP*) calculates the cosine similarity between text features and generated-image features extracted by CLIP. *BLIP-CLIP* [9] (denoted as *B-CLIP*) applies BLIP [13] to generate captions for the generated images, and then calculates the CLIP text-text cosine similarity between the generated captions and text prompts. *BLIP-VQA-naive* (denoted as *B-VQA-n*) applies BLIP VQA to ask a single question (e.g., a green bench and a red car?) with the whole prompt.

**Our proposed metrics.** *Disentangled BLIP-VQA* (denoted as *B-VQA*) is our proposed evaluation metric for attribute binding. *UniDet* is our proposed metric for 2D/3D-spatial relationships and numeracy evaluation metric.

For evaluating MLLM to serve as a metric, we assess three models, and propose MLLM-based metric for non-spatial relationship and complex compositions: *MiniGPT4* (denoted as *mGPT*), *ShareGPT4V* (denoted as *Share*), and *GPT-4V*. Due to the restricted quota for GPT-4V, we conducted evaluations on one-fifth of the images of T2I-CompBench++ (*i.e.*, 600 images for each category). To boost capability of MLLM, we apply Chain-of-Thought [77] (denoted as *-CoT*) for mGPT and ShareGPT4V.

**Human evaluation.** For human evaluation of each sub-category, we randomly select 25 prompts and generate 2 images per prompt, resulting in 300 images generated with 200 prompts per model in total. The testing set includes 300 prompts for each sub-category, resulting in 2,400 prompts in total. The prompt sampling rate for human evaluation is 8.33%. We utilize Amazon Mechanical Turk and ask three workers to score each generated-image-text pair independently based on the image-text alignment. The worker can choose a score from $\{1, 2, 3, 4, 5\}$ and we normalize the scores by dividing them by 5. We then compute the average score across all images and all workers.

### C. Comparison across Different Evaluation Metrics

**Comparions of different evaluation metrics and human evaluation.** We show the results of different evaluation metrics in Table II-V. Previous evaluation metrics, CLIP and B-CLIP, predict similar scores across different models and cannot reflect the differences between models. Our proposed metrics

* Because the DALLE 3 API does not allow specifying random seed and due to the high cost of the DALLE 3 API, we use DALLE 3 to generate 3 images for each prompt without specified seed.

TABLE II: Comparisons of evaluation metrics on attribute binding (color and shape), with scores unnormalized. **Bold** stands for the best score across 7 models in T2I-Compbench [40]. Blue represents the proposed metric for the category, and green stands for the human evaluation, applicable to the following Table III, IV and V.

| Model | Color | | | | | | | | Shape | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CLIP | B-CLIP | B-VQA-n | B-VQA | mGPT-CoT | GPT-4V | Share–CoT | Human | CLIP | B-CLIP | B-VQA-n | B-VQA | mGPT-CoT | GPT-4V | Share–CoT | Human |
| Stable v1-4 [1] | 0.3214 | 0.7454 | 0.5875 | 0.3765 | 0.7424 | 0.6324 | 0.6185 | 0.6533 | 0.3112 | 0.7077 | 0.6771 | 0.3576 | 0.7197 | 0.4466 | 0.5470 | 0.6160 |
| Stable v2 [1] | 0.3335 | 0.7616 | 0.7249 | 0.5065 | 0.7764 | 0.6717 | 0.7056 | 0.7747 | **0.3203** | 0.7191 | 0.7517 | 0.4221 | 0.7279 | 0.6212 | 0.5815 | 0.6587 |
| Composable + SD v2 [7] | 0.3178 | 0.7352 | 0.5744 | 0.4063 | 0.7524 | 0.6295 | 0.6597 | 0.6187 | 0.3092 | 0.6985 | 0.6125 | 0.3299 | 0.7124 | 0.5655 | 0.5339 | 0.5133 |
| Structured + SD v2 [8] | 0.3319 | 0.7626 | 0.7184 | 0.4990 | 0.7822 | 0.6671 | 0.7253 | 0.7867 | 0.3178 | 0.7177 | 0.7500 | 0.4218 | 0.7228 | 0.6348 | 0.5886 | 0.6413 |
| Attn-Exct + SD v2 [9] | 0.3374 | **0.7810** | 0.8362 | 0.6400 | **0.8194** | **0.7539** | **0.8110** | 0.8240 | 0.3189 | **0.7209** | 0.7723 | 0.4517 | 0.7299 | 0.5990 | 0.6066 | 0.6360 |
| GORS + SD v2 (ours) | **0.3395** | 0.7681 | **0.8471** | **0.6603** | 0.8067 | 0.7296 | 0.7994 | **0.8320** | 0.2973 | 0.7201 | **0.7937** | **0.4785** | **0.7303** | **0.6563** | **0.6329** | **0.7040** |

TABLE III: Comparisons of evaluation metrics on attribute binding (texture) and 2D-spatial relationship.

| Model | Texture | | | | | | | | 2D-Spatial | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CLIP | B-CLIP | B-VQA-n | B-VQA | mGPT-CoT | GPT-4V | Share–CoT | Human | CLIP | B-CLIP | UniDet | mGPT-CoT | GPT-4V | Share–CoT | Human |
| Stable v1-4 [1] | 0.3081 | 0.7111 | 0.6173 | 0.4156 | 0.7836 | 0.6158 | 0.5728 | 0.7227 | 0.3142 | 0.7667 | 0.1246 | 0.8338 | 0.4297 | 0.7110 | 0.3813 |
| Stable v2 [1] | 0.3185 | 0.7240 | 0.7054 | 0.4922 | 0.7851 | 0.6594 | 0.6100 | 0.7827 | 0.3206 | 0.7723 | 0.1342 | 0.8367 | 0.4450 | 0.7197 | 0.3467 |
| Composable + SD v2 [7] | 0.3092 | 0.6995 | 0.5604 | 0.3645 | 0.7588 | 0.5870 | 0.5259 | 0.6333 | 0.3001 | 0.7409 | 0.0800 | 0.8222 | 0.2200 | 0.6633 | 0.3080 |
| Structured + SD v2 [8] | 0.3167 | 0.7234 | 0.7007 | 0.4900 | 0.7806 | 0.6674 | 0.6108 | 0.7760 | 0.3201 | 0.7726 | 0.1386 | 0.8361 | 0.4617 | 0.7183 | 0.3467 |
| Attn-Exct + SD v2 [9] | 0.3171 | 0.7206 | 0.7830 | 0.5963 | 0.8062 | 0.7077 | 0.6758 | 0.8400 | 0.3213 | 0.7742 | 0.1455 | **0.8407** | 0.5103 | 0.7330 | 0.4027 |
| GORS + SD v2 (ours) | **0.3233** | **0.7315** | **0.7991** | **0.6287** | **0.8106** | **0.7266** | **0.7054** | **0.8573** | **0.3242** | **0.7854** | **0.1815** | 0.8362 | **0.4640** | **0.7373** | **0.4560** |

TABLE IV: Comparisons of evaluation metrics on 3D-spatial relationship and numeracy task.

| Model | 3D-Spatial | | | | | | | Numeracy | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CLIP | B-CLIP | UniDet | mGPT-CoT | GPT-4V | Share–CoT | Human | CLIP | B-CLIP | UniDet | mGPT-CoT | GPT-4V | Share–CoT | Human |
| Stable v1-4 [1] | 0.3001 | 0.7344 | 0.3030 | 0.8249 | 0.2880 | 0.6977 | 0.4987 | 0.3021 | 0.6598 | 0.4461 | 0.8427 | 0.3643 | 0.7423 | 0.5200 |
| Stable v2 [1] | 0.3100 | 0.7486 | 0.3230 | 0.8338 | 0.3600 | 0.7160 | 0.5040 | 0.3138 | 0.6588 | 0.4579 | 0.8364 | 0.4320 | 0.7653 | 0.5253 |
| Composable + SD v2 [7] | 0.2955 | 0.7300 | 0.2847 | 0.8253 | 0.2413 | 0.6853 | 0.4933 | 0.3028 | 0.6531 | 0.4261 | 0.8391 | 0.3842 | 0.7540 | 0.5187 |
| Structured + SD v2 [8] | 0.3093 | 0.7482 | 0.3224 | 0.8310 | 0.3653 | 0.7153 | 0.5027 | 0.3125 | 0.6588 | 0.4550 | **0.8440** | 0.4258 | 0.7637 | 0.5680 |
| Attn-Exct + SD v2 [9] | 0.3096 | 0.7517 | 0.3222 | **0.8365** | 0.3633 | 0.7213 | 0.5400 | 0.3130 | 0.6607 | 0.4767 | 0.8439 | 0.4660 | **0.7857** | 0.5613 |
| GORS + SD v2 (ours) | **0.3140** | **0.7615** | **0.3572** | 0.8292 | **0.3753** | **0.7327** | **0.5547** | **0.3152** | **0.6715** | **0.4841** | 0.8376 | **0.4740** | 0.7847 | **0.5747** |

TABLE V: Comparisons of evaluation metrics on the non-spatial relationship and complex compositions.

| Model | Non-spatial | | | | | | Complex | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CLIP | B-CLIP | mGPT-CoT | GPT-4V | Share–CoT | Human | CLIP | B-CLIP | 3-in-1 | mGPT-CoT | GPT-4V | Share–CoT | Human |
| Stable v1-4 [1] | 0.3079 | 0.7565 | 0.8170 | 0.7717 | 0.7487 | 0.9653 | 0.2876 | 0.6816 | 0.3080 | 0.8075 | 0.6453 | 0.7727 | 0.8067 |
| Stable v2 [1] | 0.3127 | 0.7609 | **0.8235** | 0.8153 | 0.7567 | 0.9827 | **0.3096** | 0.6893 | 0.3386 | 0.8094 | 0.6483 | **0.7783** | 0.8480 |
| Composable + SD v2 [7] | 0.2980 | 0.7038 | 0.7936 | 0.5030 | 0.6927 | 0.8120 | 0.3014 | 0.6638 | 0.2898 | 0.8083 | 0.5637 | 0.7487 | 0.7520 |
| Structured + SD v2 [8] | 0.3111 | 0.7614 | 0.8221 | 0.8127 | 0.7560 | 0.9773 | 0.3084 | **0.6902** | 0.3355 | 0.8076 | 0.6400 | 0.7777 | 0.8333 |
| Attn-Exct + SD v2 [9] | 0.3109 | 0.7607 | 0.8214 | 0.8243 | 0.7593 | 0.9533 | 0.2913 | 0.6875 | **0.3401** | 0.8078 | 0.6817 | 0.7763 | 0.8573 |
| GORS + SD v2 (ours) | **0.3193** | **0.7619** | 0.8172 | **0.8420** | **0.7637** | **0.9853** | 0.2973 | 0.6841 | 0.3328 | **0.8095** | **0.6850** | 0.7737 | **0.8680** |

(highlighted in blue column) show a similar trend with human evaluation (highlighted in green column).

**Human correlation of the evaluation metrics.** We calculate Kendall's tau ($\tau$) and Spearman's rho ($\rho$) to evaluate the ranking correlation between automatic evaluation and human evaluation. The human correlation results are illustrated in Table VI. The results verify the effectiveness of our proposed evaluation metrics, BLIP-VQA for attribute binding, UniDet-based metric for spatial relationships and numeracy. For MLLM evaluation, MiniGPT4 does not perform well in terms of correlation with human perception. Share-CoT and GPT-4V excel in terms of non-spatial and complex categories, followed by CLIPScore and 3-in-1 evaluation metrics. In shape attribute, they exhibit performances that are comparable to non-MLLM metrics. In color, texture, 2D/3D-spatial and numeracy categories, their performances are slightly lower than non-MLLM metrics.

**Conclusion on the evaluation metrics.** Based on the human correlation of different evaluation metrics, we draw the conclusion for the optimal evaluation metrics for each category or sub-category, and those optimal metrics are used to benchmark different text-to-image generation methods in the following subsections. (1) For attribute binding, the best evaluation metric is BLIP-VQA. (2) For 2D spatial relationships, 3D spatial relationships, and numeracy, the UniDet-based metric performs best. (3) For non-spatial relationships, the best metric is GPT-4V, the second-best metric is Share-CoT (ShareGPT4V with Chain-of-Thought), and the best non-MLLM metric is CLIP. (4) For complex compositions, the best metric is GPT-4V, the second-best metric is Share-CoT (ShareGPT4V with Chain-of-Thought), and the best non-MLLM metric is 3-in-1.

### *D. Discussion about MLLMs as a unified metric*

**Comparisons among different MLLMs.** We compare the human correlation among MiniGPT4, ShareGPT4V and GPT-4V. MiniGPT4 shows inadequate correlation with human perception†. ShareGPT4V and GPT-4V excel in terms of non-spatial and complex categories. In color, texture, 2D/3D-spatial and numeracy categories, their performances are slightly lower than non-MLLM metrics. ShareGPT4V ranks second in non-spatial relationship and complex, ranging between first to third place in attribute categories, third in 2D-spatial, third (in terms

† To facilitate comparisons of different MLLMs' capabilities, we consider the same evaluation metric as a unified entity. For instance, mGPT and mGPT-CoT are treated as a single entity (the same for Share and Share-CoT), and we compare them based on the highest human correlation values attained.

TABLE VI: The correlation between automatic evaluation metrics and human evaluation. Our proposed metrics demonstrate a significant improvement over existing metrics in terms of Kendall's $\tau$ and Spearmanr's $\rho$. **Bold** stands for the highest correlation score across all non-MLLM metrics. Red indicates the highest correlation score with MLLM metrics included.

| Metric | Attribute-color | | Attribute-shape | | Attribute-texture | | 2D Spatial rel | | 3D Spatial rel | | Numeracy | | Non-spatial rel | | Complex | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\tau(\uparrow)$ | $\rho(\uparrow)$ | $\tau(\uparrow)$ | $\rho(\uparrow)$ | $\tau(\uparrow)$ | $\rho(\uparrow)$ | $\tau(\uparrow)$ | $\rho(\uparrow)$ | $\tau(\uparrow)$ | $\rho(\uparrow)$ | $\tau(\uparrow)$ | $\rho(\uparrow)$ | $\tau(\uparrow)$ | $\rho(\uparrow)$ | $\tau(\uparrow)$ | $\rho(\uparrow)$ |
| CLIP | 0.1938 | 0.2773 | 0.0555 | 0.0821 | 0.2890 | 0.4008 | 0.2741 | 0.3548 | 0.2151 | 0.3009 | 0.1557 | 0.2187 | **0.2470** | **0.3161** | 0.0650 | 0.0847 |
| B-CLIP | 0.2674 | 0.3788 | 0.1692 | 0.2413 | 0.2999 | 0.4187 | 0.1983 | 0.2544 | 0.2042 | 0.2840 | 0.0859 | 0.1257 | 0.2342 | 0.2964 | 0.1963 | 0.2755 |
| B-VQA-n | 0.4602 | 0.6179 | 0.2280 | 0.3180 | 0.4227 | 0.5830 | - | - | - | - | - | - | - | - | - | - |
| B-VQA | **0.6297** | **0.7958** | **0.2707** | **0.3795** | **0.5177** | **0.6995** | - | - | - | - | - | - | - | - | - | - |
| UniDet | - | - | - | - | - | - | **0.4756** | **0.5136** | **0.3126** | **0.4262** | **0.4270** | **0.5311** | - | - | - | - |
| 3-in-1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | **0.2831** | **0.3853** |
| mGPT | 0.1197 | 0.1616 | 0.1282 | 0.1775 | 0.1061 | 0.1460 | 0.0208 | 0.0229 | 0.0980 | 0.1373 | 0.0647 | 0.0967 | 0.1181 | 0.1418 | 0.0066 | 0.0084 |
| mGPT-CoT | 0.3156 | 0.4151 | 0.1300 | 0.1805 | 0.3453 | 0.4664 | 0.1096 | 0.1239 | 0.0618 | 0.0760 | 0.1414 | 0.1769 | 0.1944 | 0.2137 | 0.1251 | 0.1463 |
| Share | 0.4429 | 0.5260 | 0.1410 | 0.1782 | 0.2858 | 0.3518 | 0.1716 | 0.1837 | 0.1394 | 0.1610 | 0.2908 | 0.3396 | 0.1541 | 0.1801 | 0.0629 | 0.1780 |
| Share–CoT | 0.4783 | 0.5847 | 0.2871 | 0.3659 | 0.4005 | 0.5094 | 0.2889 | 0.3182 | 0.2707 | 0.3186 | 0.4175 | 0.4891 | 0.3401 | 0.3622 | 0.3204 | 0.3649 |
| GPT-4V | 0.5242 | 0.6465 | 0.2668 | 0.3402 | 0.3944 | 0.4987 | 0.3456 | 0.4038 | 0.2560 | 0.3202 | 0.3777 | 0.4651 | 0.4756 | 0.5337 | 0.5070 | 0.5942 |

TABLE VII: Benchmarking on all categories with proposed metrics. **Bold** stands for the best score across 7 models in T2I-Compbench [40]. Red indicates the best score across 10 models in T2I-CompBench++.

| Model | Color | Shape | Texture | 2D-Spatial | 3D-Spatial | Numeracy | Non-Spatial | | | Complex | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B-VQA | B-VQA | B-VQA | UniDet | UniDet | UniDet | CLIP | GPT-4V | Share-CoT | 3-in-1 | GPT-4V | Share-CoT |
| Stable v1-4 [1] | 0.3765 | 0.3576 | 0.4156 | 0.1246 | 0.3030 | 0.4461 | 0.3079 | 0.7717 | 0.7487 | 0.3080 | 0.6453 | 0.7727 |
| Stable v2 [1] | 0.5065 | 0.4221 | 0.4922 | 0.1342 | 0.3230 | 0.4579 | 0.3127 | 0.8153 | 0.7567 | 0.3386 | 0.6483 | **0.7783** |
| Composable + SD v2 [7] | 0.4063 | 0.3299 | 0.3645 | 0.0800 | 0.2847 | 0.4261 | 0.2980 | 0.5030 | 0.6927 | 0.2898 | 0.5637 | 0.7487 |
| Structured + SD v2 [8] | 0.4990 | 0.4218 | 0.4900 | 0.1386 | 0.3224 | 0.4550 | 0.3111 | 0.8127 | 0.7560 | 0.3355 | 0.6400 | 0.7777 |
| Attn-Exct + SD v2 [9] | 0.6400 | 0.4517 | 0.5963 | 0.1455 | 0.3222 | 0.4767 | 0.3109 | 0.8243 | 0.7593 | 0.3401 | 0.6817 | 0.7763 |
| GORS-unbaised + SD v2 (ours) | 0.6414 | 0.4546 | 0.6025 | 0.1725 | 0.3300 | **0.4843** | 0.3158 | **0.8557** | **0.7650** | **0.3470** | 0.6753 | 0.7697 |
| GORS + SD v2 (ours) | **0.6603** | **0.4785** | **0.6287** | **0.1815** | **0.3572** | 0.4841 | **0.3193** | 0.8420 | 0.7637 | 0.3328 | **0.6850** | 0.7737 |
| Stable XL [34] | 0.5879 | 0.4687 | 0.5299 | 0.2133 | 0.3566 | 0.4988 | 0.3119 | 0.8500 | 0.7673 | 0.3237 | 0.7170 | 0.7817 |
| Pixart-$\alpha$-ft [30] | 0.6690 | 0.4927 | 0.6477 | 0.2064 | 0.3901 | 0.5058 | 0.3197 | 0.8620 | 0.7747 | 0.3433 | 0.7223 | 0.7823 |
| DALLE·3 [33] | 0.7785 | 0.6205 | 0.7036 | 0.2865 | 0.3744 | 0.5880 | 0.3003 | 0.9170 | 0.7853 | 0.3773 | 0.8653 | 0.7927 |

of $\tau$) and second (in terms of $\rho$) in 3D-spatial, second in numeracy. GPT-4V ranks first in non-spatial relationship and complex, second in color attribute, third in shape and texture attributes, second in 2D-spatial, second (in terms of $\tau$) and third (in terms of $\rho$) in 3D-spatial, third in numeracy.

**Comparisons of the effectiveness of Chain-of-Thought for MLLMs.** Utilizing Chain-of-thought [77] to stimulate MLLM's evaluation ability, we conduct the comparisons on MiniGPT4 and ShareGPT4V. With CoT, both MiniGPT4 and ShareGPT4V demonstrate improvements across most categories in terms of human correlation, highlighting the significance of effective prompting.

**Discussion about the stability of MLLMs' results.** We test the stability of the MLLM metric by conducting multiple executions on a consistent set of images. Specifically, we employ ShareGPT4V and GPT-4V to analyze 50 images from Stable v2 color category, repeating the process five times. Default parameters were used, with a temperature of 1.0 for GPT-4V and 0.2 for ShareGPT4V. As shown in Table VIII, our analysis consistently produces results, with variations remaining within 0.032 across the five executions for GPT-4V and 0.0273 for ShareGPT4V. The average standard deviations are 7.3235 for GPT-4V and 6.8741 for ShareGPT4V.

**Limitations of MLLMs as a metric.** Share-CoT and GPT-4V have limits despite their strong performance as evaluation metrics. They do not adhere to the prompts of grading guidelines very well. Our empirical observations show that Share-CoT tends to give less diverse ratings, regardless of the prompt provided. GPT-4V can comprehend the evaluation prompts, but it is less able to convert into exact grades.

### E. Benchmarking on Different Methods

The quantitative benchmarking results are reported in Table II-V. Qualitative results are shown in appendix Figure 5, Figure 6 and Figure 7.

**Comparisons across text-to-image models.** (1) Stable Diffusion v2 consistently outperforms Stable Diffusion v1-4 in all types of compositional prompts and evaluation metrics. (2) Although Structured Diffusion built upon Stable Diffusion v1-4 shows great performance improvement in attribute binding as reported in Feng *et al.* [8], Structured Diffusion built upon Stable Diffusion v2 only brings slight performance gain upon Stable Diffusion v2. It indicates that boosting the performance upon a better baseline of Stable Diffusion v2 is more challenging. (3) Composable Diffusion built upon Stable Diffusion v2 does not work well. A similar phenomenon was also observed in previous work [9] that Composable Diffusion often generates images containing a mixture of the subjects. In addition, Composable Diffusion was designed for concept conjunctions and negations so it is reasonable that it does not perform well in other compositional scenarios. (4) Attend-and-Excite built upon Stable Diffusion v2 improves the performance in attribute binding. (5) Previous methods Composable Diffusion [7], Structure Diffusion [8] and Attend-and-Excite [9] are designed for concept conjunction or attribute binding, so they do not result in significant improvements in object relationships. (6) Our proposed approach, GORS, outperforms previous approaches across all types of compositional prompts,

TABLE VIII: Stability tests on MLLM metrics

| Test round | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Share–CoT | 0.7232 | 0.7109 | 0.7267 | 0.6959 | 0.7129 |
| GPT4V | 0.6918 | 0.7078 | 0.6998 | 0.6758 | 0.6853 |

as demonstrated by the automatic evaluation, human evaluation, and qualitative results. The evaluation results of GORS-unbiased and GORS significantly exceed the baseline Stable v2. Besides, the performances of GORS-unbiased indicate that our proposed approach is insensitive to the reward model used for selecting samples, and that the proposed approach works well as long as high-quality samples are selected. (7) Recent text-to-image models, Stable Diffusion XL [34], Pixart-$\alpha$-ft [30] and DALLE 3 [33] improves across all categories compared to previous methods. Notably, DALLE 3 achieves the state-of-the-art (SOTA) performance by a significant margin in almost all categories.

**Comparisons across compositionality categories.** According to the human evaluation results, spatial relationship is the most challenging sub-category for text-to-image models, and attribute binding (shape) is also challenging. Non-spatial relationship is the easiest sub-category.

**Comparisons between seen and unseen splits.** We provide the seen and unseen splits for the test set in attribute binding, where the unseen set consists of attribute-object pairs that do not appear in the training set. The unseen split tends to include more uncommon attribute-object combinations than seen split. The performance comparison of seen and unseen splits for attribute binding is shown in Table XVI. Our observations reveal that our model exhibits slightly lower performance on the unseen set than the seen set.

### F. Ablation Study

We conduct ablation studies on our proposed GORS approach and evaluation metric.

**Finetuning strategy.** Our approach finetunes both the CLIP text encoder and the U-Net of Stable Diffusion with LoRA [81]. We conduct with the attribute binding (color) sub-category. We investigate the effects of finetuning CLIP only and U-Net only with LoRA. As shown in Table IX, our model which finetunes both CLIP and U-Net performs better.

**Threshold of selecting samples for finetuning.** Our approach fine-tunes Stable Diffusion v2 with the selected samples that align well with the compositional prompts. We manually set a threshold for the alignment score to select samples with higher alignment scores than the threshold for fine-tuning. We experiment with setting the threshold to half of its original value, and setting the threshold to 0 (*i.e.*, use all generated images for finetuning with rewards, without selection). Results in Table IX demonstrate that half threshold and zero threshold will lead to worse performance.

**Qualitative results of ablation study** We show the qualitative results of the variants in ablation study in Figure 4. When only CLIP is fine-tuned with LoRA, the generated images do not bind attributes to correct objects (for example, Figure 4 Row. 3 Col. 3 and Row. 6 Col. 3). Noticeable improvements are observed in the generated images when U-Net is fine-tuned by LoRA, particularly when both CLIP and U-Net are finetuned together. Furthermore, we delve into the effect of the threshold for selecting images aligned with text prompts for fine-tuning. A higher threshold value enables the selection of images that are highly aligned with text prompts for finetuning, ensuring that only well-aligned examples are incorporated into the finetuning process. In contrast, a lower threshold leads to the inclusion of misaligned images during finetuning, which can degrade the compositional ability of the finetuned text-to-image models (for example, Figure 4 last two columns in Row. 2).

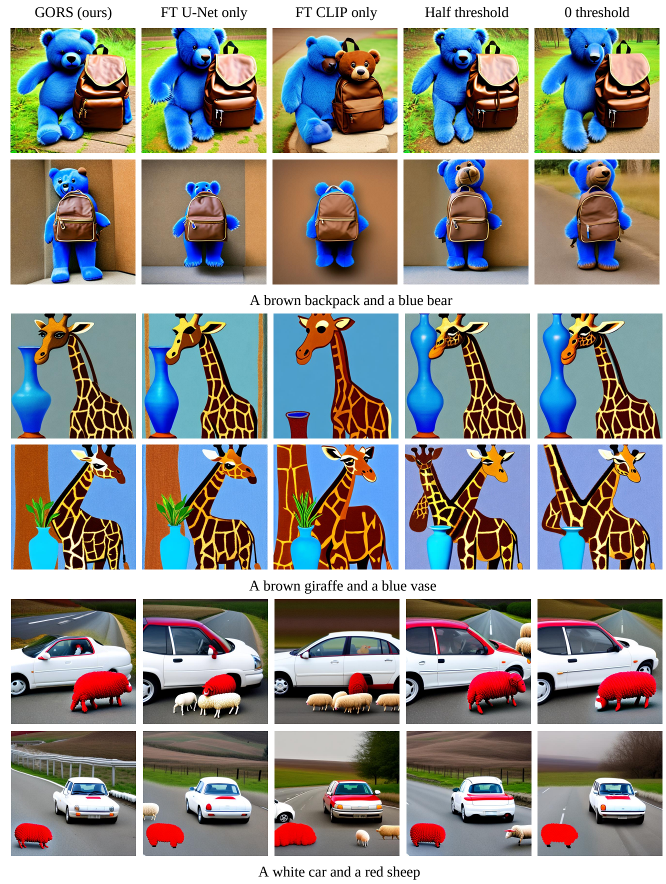


Fig. 4: Qualitative comparison of ablation study on fine-tuning strategy and threshold.

## VII. Conclusion and Discussions

We propose T2I-CompBench++, a comprehensive benchmark for open-world compositional text-to-image generation, consisting of 8,000 prompts from 4 categories and 8 sub-categories. We propose new evaluation metrics and an improved baseline for the benchmark, and validate the effectiveness of the metrics and method by extensive evaluation. We further study the potential and limitations of MLLMs as a unified metrics. Besides, we propose a simple yet effective way GORS to boost the compositionally of text-to-image

TABLE IX: Ablation studies on fine-tuning strategy and threshold.

| Metric | FT U-Net only | FT CLIP only | Half threshold | 0 threshold | GORS (ours) |
|---|---|---|---|---|---|
| B-VQA | 0.6216 | 0.5507 | 0.6157 | 0.6130 | **0.6570** |

models. When studying generative models, researchers need to be aware of the potential negative social impact, for example, it might be abused to generate fake news. We also need to be aware of the bias from the image generators and the evaluation metrics based on pretrained multimodal models.

## REFERENCES


[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.
[2] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *The Journal of Machine Learning Research*, 2022.
[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *NeurIPS*, 2022.
[4] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021.
[5] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *ICLR*, 2021.
[6] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein *et al.*, "Muse: Text-to-image generation via masked generative transformers," *arXiv preprint arXiv:2301.00704*, 2023.
[7] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *ECCV*, 2022.
[8] W. Feng, X. He, T.-J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, "Training-free structured diffusion guidance for compositional text-to-image synthesis," in *ICLR*, 2023.
[9] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," in *ACM Trans. Graph.*, 2023.
[10] Q. Wu, Y. Liu, H. Zhao, T. Bui, Z. Lin, Y. Zhang, and S. Chang, "Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis," in *ICCV*, 2023.
[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
[12] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.
[13] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.
[14] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.
[15] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.
[16] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "Sharegpt4v: Improving large multi-modal models with better captions," *arXiv preprint arXiv:2311.12793*, 2023.
[17] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, vol. 9, no. 1, p. 1, 2023.
[18] E. M. Bakr, P. Sun, X. Shen, F. F. Khan, L. E. Li, and M. Elhoseiny, "Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models," in *ICCV*, 2023.
[19] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *ICML*, 2016.
[20] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *NeurIPS*, 2016.
[21] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.
[22] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018.
[23] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *CVPR*, 2019.
[24] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *CVPR*, 2021.
[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.
[26] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.
[27] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *ICML*, 2022.
[28] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *NeurIPS*, 2022.
[29] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman, "Make-a-scene: Scene-based text-to-image generation with human priors," in *ECCV*, 2022.
[30] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu *et al.*, "Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis," in *ICLR*, 2024.
[31] J. Gu, S. Zhai, Y. Zhang, J. M. Susskind, and N. Jaitly, "Matryoshka diffusion models," 2023.
[32] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021.
[33] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.
[34] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.
[35] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," 2024.
[36] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, N. Yu, Z. Chen, H. Wang, S. Savarese, S. Ermon *et al.*, "Hive: Harnessing human feedback for instructional visual editing," *arXiv preprint arXiv:2303.09618*, 2023.
[37] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu, "Aligning text-to-image models using human feedback," *arXiv preprint arXiv:2302.12192*, 2023.
[38] H. Dong, W. Xiong, D. Goyal, R. Pan, S. Diao, J. Zhang, K. Shum, and T. Zhang, "Raft: Reward ranked finetuning for generative foundation model alignment," *arXiv preprint arXiv:2304.06767*, 2023.
[39] Z. Li, M. R. Min, K. Li, and C. Xu, "Stylet2i: Toward compositional and high-fidelity text-to-image synthesis," in *CVPR*, 2022.
[40] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *NeurIPS*, 2024.
[41] M. Patel, C. Kim, S. Cheng, C. Baral, and Y. Yang, "Eclipse: A resource-efficient text-to-image prior for image generations," *arXiv preprint arXiv:2312.04655*, 2023.
[42] X. Liu, T. Hu, W. Wang, K. Kawaguchi, and Y. Yao, "Referee can play: An alternative approach to conditional generation via model inversion," *arXiv preprint arXiv:2402.16305*, 2024.
[43] D. H. Park, S. Azadi, X. Liu, T. Darrell, and A. Rohrbach, "Benchmark for compositional text-to-image synthesis," in *NeurIPS*, 2021.
[44] L. Lian, B. Li, A. Yala, and T. Darrell, "Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models," *arXiv preprint arXiv:2305.13655*, 2023.

[45] M. Chen, I. Laina, and A. Vedaldi, "Training-free layout control with cross-attention guidance," in *WACV*, 2024.
[46] R. Wang, Z. Chen, C. Chen, J. Ma, H. Lu, and X. Lin, "Compositional text-to-image synthesis with attention map control of diffusion models," *arXiv preprint arXiv:2305.13921*, 2023.
[47] T. H. S. Meral, E. Simsar, F. Tombari, and P. Yanardag, "Conform: Contrast is all you need for high-fidelity text-to-image diffusion models," *arXiv preprint arXiv:2312.06059*, 2023.
[48] Y. Kim, J. Lee, J.-H. Kim, J.-W. Ha, and J.-Y. Zhu, "Dense text-to-image generation with attention modulation," 2023.
[49] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik, "Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment," *NeurIPS*, 2024.
[50] H. Gani, S. F. Bhat, M. Naseer, S. Khan, and P. Wonka, "Llm blueprint: Enabling text-to-image generation with complex and detailed prompts," *arXiv preprint arXiv:2310.10640*, 2023.
[51] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *ICCV*, 2023.
[52] A. Taghipour, M. Ghahremani, M. Bennamoun, A. M. Rekavandi, H. Laga, and F. Boussaid, "Box it to bind it: Unified layout control and attribute binding in t2i diffusion models," *arXiv preprint arXiv:2402.17910*, 2024.
[53] Z. Wang, E. Xie, A. Li, Z. Wang, X. Liu, and Z. Li, "Divide and conquer: Language models can plan and self-correct for compositional text-to-image generation," *arXiv preprint arXiv:2401.15688*, 2024.
[54] X. Chen, Y. Liu, Y. Yang, J. Yuan, Q. You, L.-P. Liu, and H. Yang, "Reason out your layout: Evoking the layout master from large language models for text-to-image synthesis," *arXiv preprint arXiv:2311.17126*, 2023.
[55] L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon, and B. Cui, "Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms," *arXiv preprint arXiv:2401.11708*, 2024.
[56] X. Zhang, L. Yang, Y. Cai, Z. Yu, J. Xie, Y. Tian, M. Xu, Y. Tang, Y. Yang, and B. Cui, "Realcompo: Dynamic equilibrium between realism and compositionality improves text-to-image diffusion models," *arXiv preprint arXiv:2402.12908*, 2024.
[57] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.
[58] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *ICVGIP*, 2008.
[59] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[60] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models," in *ICCV*, 2023.
[61] V. Petsiuk, A. E. Siemenn, S. Surbehera, Z. Chin, K. Tyser, G. Hunter, A. Raghavan, Y. Hicke, B. A. Plummer, O. Kerret *et al.*, "Human evaluation of text-to-image models on a multi-task benchmark," *arXiv preprint arXiv:2211.12112*, 2022.
[62] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NeurIPS*, 2016.
[63] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017.
[64] Y. Lu, X. Yang, X. Li, X. E. Wang, and W. Y. Wang, "Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation," *arXiv preprint arXiv:2305.11116*, 2023.
[65] Y. Chen, "X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models," *arXiv preprint arXiv:2305.10843*, 2023.
[66] S. Wen, G. Fang, R. Zhang, P. Gao, H. Dong, and D. Metaxas, "Improving compositional text-to-image generation with large vision-language models," *arXiv preprint arXiv:2310.06311*, 2023.
[67] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," 2023.
[68] J. Sun, D. Fu, Y. Hu, S. Wang, R. Rassin, D.-C. Juan, D. Alon, C. Herrmann, S. van Steenkiste, R. Krishna *et al.*, "Dreamsync: Aligning text-to-image generation with image understanding feedback," *arXiv preprint arXiv:2311.17946*, 2023.
[69] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *NeurIPS*, 2024.
[70] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Better aligning text-to-image models with human preference," 2023.
[71] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," *arXiv preprint arXiv:2306.09341*, 2023.
[72] Y. Liang, J. He, G. Li, P. Li, A. Klimovskiy, N. Carolan, J. Sun, J. Pont-Tuset, S. Young, F. Yang *et al.*, "Rich human feedback for text-to-image generation," *arXiv preprint arXiv:2312.10240*, 2023.
[73] M. Ku, T. Li, K. Zhang, Y. Lu, X. Fu, W. Zhuang, and W. Chen, "Imagenhub: Standardizing the evaluation of conditional image generation models," *arXiv preprint arXiv:2310.01596*, 2023.
[74] M. Ku, D. Jiang, C. Wei, X. Yue, and W. Chen, "Viescore: Towards explainable metrics for conditional image synthesis evaluation," *arXiv preprint arXiv:2312.14867*, 2023.
[75] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente *et al.*, "Holistic evaluation of text-to-image models," *NeurIPS*, 2024.
[76] OpenAI, https://openai.com/blog/chatgpt/, 2023.
[77] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.
[78] X. Zhou, V. Koltun, and P. Krähenbühl, "Simple multi-dataset detection," in *CVPR*, 2022.
[79] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *ICCV*, 2021.
[80] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[81] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
[82] D. Teney, E. Abbasnejad, K. Kafle, R. Shrestha, C. Kanan, and A. Van Den Hengel, "On the value of out-of-distribution testing: An example of goodhart's law," *Advances in neural information processing systems*, vol. 33, pp. 407–417, 2020.
[83] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
[84] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *CVPR*, 2022.
[85] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *ICCV*, 2019.
[86] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," in *IJCV*, 2020.
[87] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017.
[88] https://github.com/huggingface/diffusers/blob/main/examples/text_to_image.
[89] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2018.

# Appendix

**Prompt template for MLLM evaluation.** Specifically, Table XI shows the prompts for evaluating attribute binding (color, shape, texture). Similar to BLIP-VQA, for each noun phrase in a prompt, we request a number equivalent to the noun phrase and perform a multiplication. Table XII, Table XIV, Table XIII, and Table XV demonstrate the prompt templates used for 2D/3D-spatial relationships, non-spatial relationships, numeracy, and complex compositions, respectively[‡]. Due to the robust capabilities of GPT-4V and limited quota constraints, we have omitted "describe the image" for prompt of GPT-4V.

**Comparisons of the scalability of our proposed approach.** To demonstrate the scalability of our proposed approach, we introduce additional 700 prompts of complex compositions to form an extended training set of 1,400 complex prompts. The new prompts are generated with the same methodology . We conduct 6 experiments to train the models with different training set sizes, i.e., 25 prompts, 275 prompts, 350 prompts, 700 prompts, 1050 prompts, and 1400 prompts. The results in Table XVII show the performance of our model grows with the increase of the training set sizes. The results indicate the potential to achieve better performance by scaling up the training set.

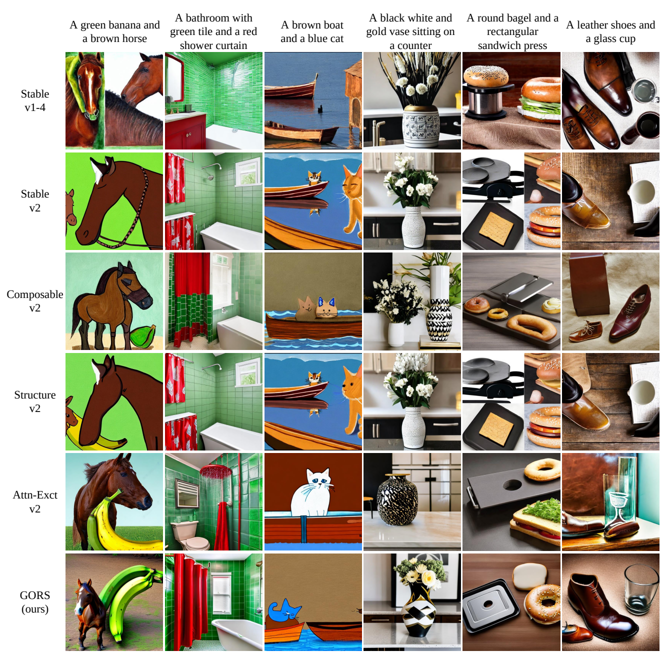


Fig. 5: Qualitative comparison between our approach and previous methods.

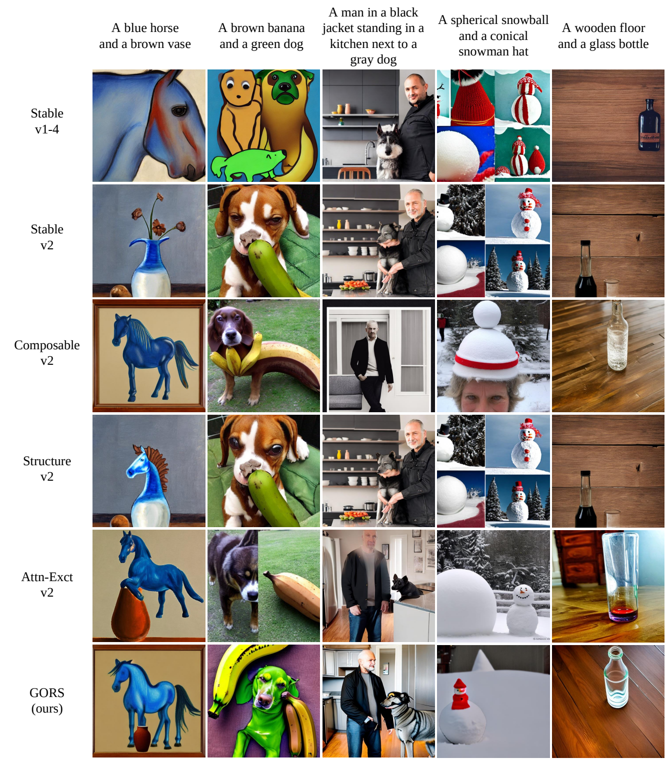


Fig. 6: Qualitative comparison between our approach and previous methods.

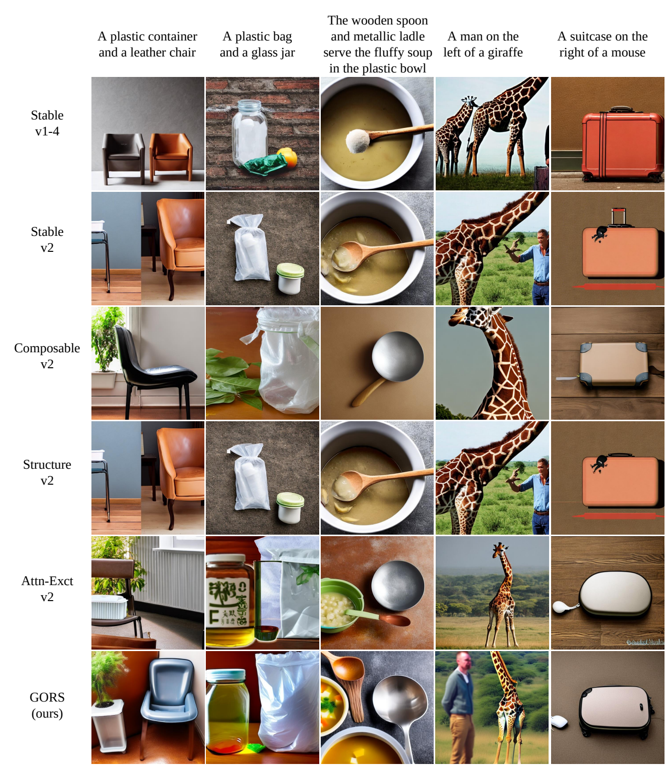


Fig. 7: Qualitative comparison between our approach and previous methods.

‡ We empirically observed that, using a five-graded marking system instead of hundred-mark system enhances the performances of ShareGPT4V.

TABLE X: Textural attributes and associated objects to construct the attribute-texture prompts.

| Textures | Objects |
|---|---|
| Rubber | band, ball, tire, gloves, sole shoes, eraser, boots, mat |
| Plastic | Bottle, bag, toy, cutlery, chair, phone case, container, cup, plate |
| Metallic | car, jewelry, watch, keychain, desk lamp, door knob, spoon, fork, knife, key, ring, necklace, bracelet, earring |
| Wooden | chair, table, picture frame, toy, jewelry box, door, floor, chopsticks, pencils, spoon, knife |
| Fabric | bag, pillow, curtain, shirt, pants, dress, blanket, towel, rug, hat, scarf, sweater, jacket |
| Fluffy | pillow, blanket, teddy bear, rug, sweater, clouds, towel, scarf, hat |
| Leather | jacket, shoes, belt, bag, wallet, gloves, chair, sofa, hat, watch |
| Glass | bottle, vase, window, cup, mirror, jar, table, bowl, plate |

TABLE XI: Prompts details for MLLM evaluation on attribute binding.

| | |
|---|---|
| Describe | You are my assistant to identify any objects and their color (shape, texture) in the image.<br>Briefly describe what it is in the image within 50 words. |
| Predict | According to the image and your previous answer, evaluate if there is {adj.+noun} in the image.<br>Give a score from 0 to 100, according the criteria:<br>100: there is {noun}, and {noun} is {adj}.<br>75: there is {noun}, {noun} is mostly {adj}.<br>20: there is {noun}, but it is not {adj}.<br>10: no {noun} in the image.<br>Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words). |

TABLE XII: Prompts details for MLLM evaluation on 2D/3D-spatial relationship.

| | |
|---|---|
| Describe | You are my assistant to identify objects and their spatial layout in the image.<br>Briefly describe the image within 50 words. |
| Predict | According to the image and your previous answer, evaluate if the text "{xxx}" is correctly portrayed in the image.<br>Give a score from 0 to 100, according the criteria:<br>100: correct spatial layout in the image for all objects mentioned in the text.<br>80: basically, spatial layout of objects matches the text.<br>60: spatial layout not aligned properly with the text.<br>40: image not aligned properly with the text.<br>20: image almost irrelevant to the text.<br>Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words). |

TABLE XIII: Prompts details for MLLM evaluation on non-spatial relationship.

| | |
|---|---|
| Describe | You are my assistant to identify the actions, events, objects and their relationships in the image.<br>Briefly describe the image within 50 words. |
| Predict | According to the image and your previous answer, evaluate if the text "{xxx}" is correctly portrayed in the image.<br>Give a score from 0 to 100, according the criteria:<br>100: the image accurately portrayed the actions, events and relationships between objects described in the text.<br>80: the image portrayed most of the actions, events and relationships but with minor discrepancies.<br>60: the image depicted some elements, but action relationships between objects are not correct.<br>40: the image failed to convey the full scope of the text.<br>20: the image did not depict any actions or events that match the text.<br>Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words). |

TABLE XIV: Prompts details for MLLM evaluation on numeracy.

| | |
|---|---|
| Describe | You are my assistant to identify objects and their quantities in the image.<br>Briefly describe the image within 50 words. |
| Predict | According to the image and your previous answer, evaluate how well the image aligns with the text prompt: "{xxx}".<br>Give a score from 0 to 100, according the criteria:<br>100: correct numerical content in the image for all objects mentioned in the text.<br>80: basically, numerical content of objects matches the text.<br>60: numerical content not aligned properly with the text.<br>40: image not aligned properly with the text.<br>20: image almost irrelevant to the text.<br>Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words). |

TABLE XV: Prompts details for MLLM evaluation on complex compositions.

| | |
|---|---|
| Describe | You are my assistant to evaluate the correspondence of the image to a given text prompt.<br>Briefly describe the image within 50 words, focus on the objects in the image and their attributes (such as color, shape, texture), spatial layout and action relationships. |
| Predict | According to the image and your previous answer, evaluate how well the image aligns with the text prompt: {xxx}.<br>Give a score from 0 to 100, according the criteria:<br>100: the image perfectly matches the content of the text prompt, with no discrepancies.<br>80: the image portrayed most of the actions, events and relationships but with minor discrepancies.<br>60: the image depicted some elements in the text prompt, but ignored some key parts or details.<br>40: the image did not depict any actions or events that match the text.<br>20: the image failed to convey the full scope in the text prompt.<br>Provide your analysis and explanation in JSON format with the following keys: score (e.g., 85), explanation (within 20 words). |

TABLE XVI: Performances of our model on attribute binding (color, shape, and texture) for seen and unseen sets.

| **Metric** | Color | | Shape | | Texture | |
|---|---|---|---|---|---|---|
| | Seen | unseen | Seen | unseen | Seen | unseen |
| CLIP | **0.3422** | 0.3283 | 0.2926 | **0.3068** | **0.3240** | 0.3219 |
| B-CLIP | **0.7716** | 0.7612 | **0.7425** | 0.6752 | **0.7569** | 0.6809 |
| B-VQA | **0.7192** | 0.5426 | **0.5500** | 0.3356 | **0.7647** | 0.3567 |
| mGPT | **0.6626** | 0.6780 | **0.6381** | 0.6307 | **0.6773** | 0.6580 |
| mGPT-CoT | **0.8082** | 0.8038 | **0.7510** | 0.6888 | **0.8453** | 0.7412 |
| Share | **0.8337** | 0.7657 | 0.6014 | **0.6219** | **0.6875** | 0.5084 |
| Share-CoT | **0.7938** | 0.7634 | **0.6692** | 0.6222 | **0.7641** | 0.6400 |
| GPT4V | **0.7491** | 0.6908 | **0.6824** | 0.6042 | **0.7916** | 0.5968 |

TABLE XVII: Performances of our model on complex compositons on the 3-in-1 metric

| ours (25) | ours (275) | ours (350) | ours (700) | ours (1050) | ours (1400) |
|---|---|---|---|---|---|
| 0.2596 | 0.3086 | 0.3299 | 0.3328 | 0.3371 | **0.3504** |