ConceptMix: A Compositional Image Generation Benchmark with Controllable Difficulty

Xindi Wu[*]   Dingli Yu[*]   Yangsibo Huang[*]   Olga Russakovsky   Sanjeev Arora

Princeton University

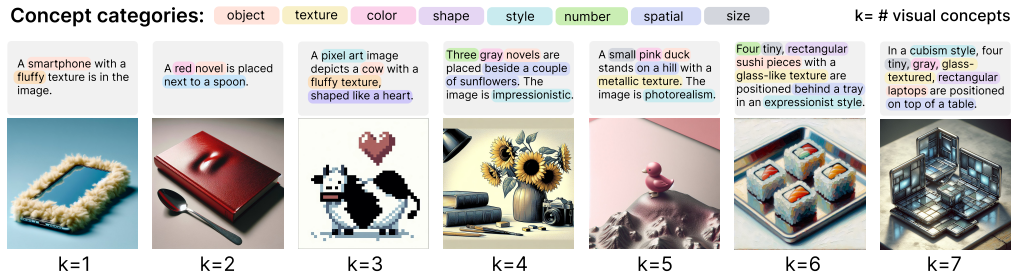



Figure 1: **Overview of our CONCEPTMIX benchmark.** CONCEPTMIX evaluates compositional generation capability of Text-to-Image (T2I) models. We show several images generated by DALL·E 3 [2] with different levels of compositional complexity k ($k = 1...7$, k denotes number of additional visual concepts other than the default object, $k = 0$ means one object, $k = 1$ means an object with one additional concept). Given text prompts with k randomly sampled visual concepts, CONCEPTMIX provides a scalable, controllable and customizable benchmark for compositional T2I evaluation.

Abstract

Compositionality is a critical capability in Text-to-Image (T2I) models, as it reflects their ability to understand and combine multiple concepts from text descriptions. Existing evaluations of compositional capability rely heavily on human-designed text prompts or fixed templates, limiting their diversity and complexity, and so the evaluations have low discriminative power. We propose CONCEPTMIX, a scalable, controllable, and customizable benchmark consisting of two stages: (a) With categories of visual concepts (e.g., objects, colors, shapes, spatial relationships), it *randomly* samples an object and k-tuples of visual concepts to generate text prompts with GPT-4o for image generation. (b) To automatically evaluate generation quality, CONCEPTMIX uses an LLM to generate one question per visual concept, allowing automatic grading of whether each specified concept appears correctly in the generated images. By testing a diverse set of T2I models using increasing values of k, we show that our CONCEPTMIX has higher discrimination power than earlier benchmarks. CONCEPTMIX reveals, unlike previous benchmarks, the performance of several models drops dramatically with increased k. CONCEPTMIX is easily extendable to more visual concept categories and gives insight into lack of prompt diversity in datasets such as LAION-5B, guiding future T2I model development.

---

[*]Equal contribution

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Table 1: **Comparison of Compositional T2I Benchmarks.** Unlike prior benchmarks that rely on fixed templates with restricted concept categories and a constrained number of concepts per prompt, which limits the evaluation of a model's compositional generation capability, our CONCEPTMIX offers a flexible, GPT-4o-driven approach, supporting **all possible combinations** of concepts and an unlimited number of concepts in each prompt.

Benchmark	Concept Diversity	Concept Binding Method	# Concepts in Each Text Prompt
CC-500 [12]	2 categories	Fixed template	2
ABC-6K [12]	2 categories	Fixed template	2
Attn-Exct [6]	4 categories	Fixed template	2
HRS-comp [1]	2 categories	Fixed template	≤ 3
T2I-CompBench [19]	6 categories	Fixed template, ChatGPT augmented	≤ 5
CONCEPTMIX (ours)	8 categories	**Free-form**, GPT-4o generated	Unlimited

1 Introduction

Visual concepts form the building blocks of compositional Text-to-Image (T2I) generation. T2I generation has made remarkable progress [35, 43, 26, 33] with the rise of diffusion models [42, 17]. However, even top-performing models still struggle with generating images from complex prompts involving multiple *visual concepts*, such as numbers, colors, and spatial relationships. Moreover, evaluating these generated results remain challenging. Traditional perceptual metrics (e.g. FID [15], IS [38], LPIPS [48]) and embedding based approaches (e.g. CLIP [34]) often fail to capture the fine-grained text-image misalignments, such as whether the dog is standing in front of or behind the cat in an image. Such limitations of perceptual metrics become more problematic when measuring the compositional capability of T2I models with an increasing number of *visual concepts*.

Why is Compositional T2I Evaluation hard? Despite many existing benchmarks focusing on compositionality [19, 28], developing a comprehensive and expandable compositional T2I benchmark is particularly challenging for several reasons. First, existing benchmarks often cover only a subset of *visual concepts* due to limitations in prompt creation. Second, most evaluations lack scalability and flexibility, typically capping at five concepts per prompt due to the fixed templates for concept combination (e.g., "a {adj} {noun}"). This makes it hard to adapt towards more complex evaluations. In Tab. 1, we summarize the diversity and complexity of visual concepts and their composition in existing compositional benchmarks.

CONCEPTMIX. In this work, we propose CONCEPTMIX, a scalable and flexible benchmark that evaluates the compositional generation capabilities of T2I models. CONCEPTMIX uses GPT-4o [31] to create prompt by combining one random object with k random visual concepts without fixed templates. Concretely, we consider eight categories of visual concepts, including objects, colors, numbers, shapes, sizes, textures, styles, and spatial relationships. The resulting prompts of CONCEPT-MIX are much more diverse and complex compared to existing benchmarks, especially when k is large. Our prompt generation pipeline also enables efficient and accurate prompt decomposition, thus we can evaluate results base on each individual concept and aggregate the results as the final score for each image. Fig. 2 provides an overview of CONCEPTMIX along with a $k = 4$ example.

Our prompt generation is partly inspired by SKILL-MIX [47], a recent evaluation that measures the capability of large language models (LLMs) to generate a short piece of text exhibiting a random subset of language skills under a random topic. Like SKILL-MIX, our prompt generation allows easy updating and expansion of the visual concepts to be evaluated, which is demonstrated later in §3.3 where we create variants of CONCEPTMIX. Additionally, the number of possible combinations of visual concepts grows exponentially with k. Thus, with a large k, CONCEPTMIX can generate millions of unique prompts, making it impossible for models to cheat by simply memorizing or overfitting to its training set. In consequence, CONCEPTMIX offers a precise and discriminative approach to identify differences in capabilities that may not be captured by traditional leaderboards or benchmarks. This provides a better understanding of a model's strengths and weaknesses and encourages the development of models that can combine visual concepts in meaningful and creative ways. We summarize our **main contributions** as follows:

2

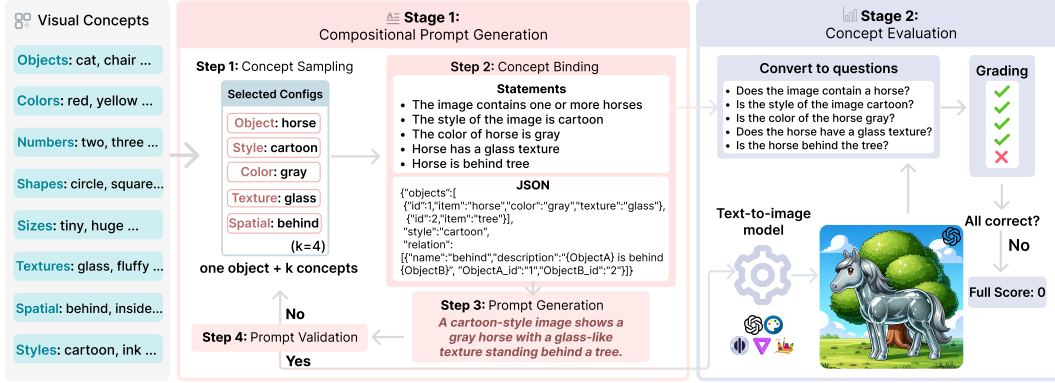



Figure 2: **CONCEPTMIX**. CONCEPTMIX consists of two main stages: 1) **Compositional Prompt Generation**: We randomly select visual concepts from 8 categories and combine them to form generation statements and intermediate JSON files with GPT-4o assistance. The statements and JSON structure are then used by GPT-4o to generate a text prompt, which, if valid, is fed into a T2I model to produce an image. 2) **Concept Evaluation**: The generated image is graded based on how well it matches with each visual concepts. This is done by converted the generation statements into questions and evaluating the answers. The image receives a score of 1 if it correctly matches all concepts, and 0 if any concept is not satisfied.

1. We introduce CONCEPTMIX (§2), the first T2I benchmark capable of evaluating the compositional generation with more than five visual concepts. By dynamically combining concepts from eight different categories, CONCEPTMIX can generate a vast set of unique prompts, evaluating a model's ability to generalize beyond its training data.

2. Our systematic evaluation of eight state-of-the-art T2I models reveals a consistent performance drop as k increases, showing the difficulty level of CONCEPTMIX can be easily controlled by k (§3.3). Even the leading proprietary model, DALL·E 3, generates full-mark[2] images for only 17% of text prompts on CONCEPTMIX with $k = 5$.

3. CONCEPTMIX clearly differentiates T2I models compared to previous compositional benchmarks [19], especially with $k \geq 2$ (§3.4). It also provides customizable evaluation by accommodating concept difficulty disparities (§3.2), resulting in easy and hard variants of CONCEPTMIX.

4. Most models' chances of generating full-mark images drop below 25% at $k = 3$ and below 10% at $k = 4$ (Tab. 3). We trace this performance limitation to training corpora like LAION [40], which we find to severely lack complex visual concept combinations beyond $k = 3$ (§3.5).

Our study highlights the pressing need for more challenging benchmarks to better differentiate T2I model performance and identify their limitations in compositional generation. Moreover, our findings highlight the critical need for better training data with diverse and complex visual concept combinations to improve the compositional generation capabilities of T2I models.

2 ConceptMix

2.1 Overview

CONCEPTMIX evaluates T2I models' ability to compose k randomly chosen visual concepts, where k controls the difficulty level. CONCEPTMIX categorizes visually interpretable concepts into eight categories, including objects, colors, numbers, and spatial relationships, etc. We define difficulty level k as the number of *extra* concepts added to an image beyond the default object[3], resulting in CONCEPTMIX(k). For example, CONCEPTMIX(1) evaluates a model's ability to generate images containing a random object and another random visual concept. Since CONCEPTMIX(0) involves no compositionality, we focus on $k \geq 1$ for the rest of the paper.

We carefully design CONCEPTMIX with two main objectives: 1) generating coherent text prompts from randomly selected concepts, and 2) automatically grading images based on complex prompts, particularly as the difficulty level (k) increases. To tackle the first goal, we carefully select the sets of concepts (§2.2) and designing a four-step pipeline for generating and validating the text prompts

---

[2]Full mark means the image correctly composes the given object and all k visual concepts.

[3]This is reasonable as image captions usually contain at least one object as the noun.

3

Table 2: **Concept Categories in CONCEPTMIX.** We collect eight diverse visual concept categories in CONCEPTMIX to cover a wide range of visual concepts commonly used in compositional T2I generation. For each category, we provide definition, concepts, and appearances in our text prompts.

Category	Concepts	Appearances in Text Prompts
Objects	car, chair, sushi, etc.	A **woman** is holding a **ring** in her hand
Colors	red, yellow, pink, etc.	A single **blue** dog is present in the image.
Numbers	two, three, four, etc.	The image shows exactly **four** sheep standing on a grassy field.
Shapes	circle, square, triangle, etc.	An oak tree with a **heart-shaped** outline stands prominently in the scene.
Sizes	tiny, huge, etc.	A **huge** cow is standing next to a sheep.
Textures	metallic, glass, fluffy, etc.	The image features a house with a **glass texture**.
Spatial	on top of, behind, inside, etc.	The image shows a bench with an oak tree positioned **behind** it
Styles	cartoon, sketch, watercolor, etc.	A **sketch** shows a single ring drawn with simple lines.

88 (§2.3). Building on this pipeline, we develop evaluation methods in §2.4 to grade the presence of the
89 required concepts in the generated images and to aggregate a final evaluation score.

2.2 Selecting Visual Concepts

91 CONCEPTMIX includes eight categories of visual concepts: objects, colors, numbers, textures, shapes,
92 sizes, styles, and spatial relationships, covering a much wider range of concepts than prior work
93 [19] (see Tab. 2 for descriptions and examples). To ensure valid text prompts[4], we exclude concept
94 categories where eligibility heavily depends on the object (e.g., actions)[5]. This exclusion is important
95 because our selection of concepts is random, and even though we have a filtering mechanism in the
96 pipeline (see §2.3), including categories like actions would still harm the efficiency of evaluation.

97 For each category, we identify representative concepts from existing literature [19, 28] and supplement
98 them with a diverse set generated by GPT-4. We then filter concepts that: 1) rarely combine with
99 others (e.g., "spongy" texture), 2) are challenging for current T2I models even individually [44] (e.g.,
100 the number "6"), and 3) are difficult to judge objectively (e.g., "median" size, "minimalism" style).

2.3 Compositional Prompt Generation

102 CONCEPTMIX(k) evaluates compositional capability by randomly sampling one object and k con-
103 cepts, and prompting T2I models to generate images containing all of them. This process involves four
104 steps: 1) randomly select k concept categories and choose concepts from them (**concept sampling**),
105 2) generate a description for each concept and create a JSON representation of the binding structure
106 (**concept binding**), 3) generate a text prompt based on the binding structure (**prompt generation**),
107 and 4) validate the generated text prompt using GPT-4o (**prompt validation**). Details of each step
108 and the GPT-4o query templates are provided in Appendix B.

109 **Step 1: Concept Sampling.** We first sample the concept categories for the $k + 1$ concepts, then
110 sample specific concepts in corresponding categories. We always ensure that the first concept is an
111 object. The remaining k concepts have a $1/4$ chance of being objects and a $3/4$ chance of being
112 sampled from the other seven categories. We resample if there is more than one concept from the style
113 category or if the number of concepts from any category (except for the spatial category) exceeds the
114 number of objects.

115 **Step 2: Concept Binding.** For concepts from the color, number, shape, size, or texture categories,
116 we randomly select an object and bind the concept to it. If spatial is selected as one of the k categories,
117 we ask GPT-4o to bind each spatial concept with two objects.[6] In some cases, a concept may need a
118 reference object to be accurately illustrated. For example, one cannot judge if an object is tiny or
119 not if it is the only object in the image. In such cases, we also request GPT-4o to add appropriate
120 reference objects. We formalize the binding as $k + 1$ statements (one for each concept) and a JSON
121 object. In Fig. 2, we provide an example ($k = 4$) demonstrating the concept binding process.

---

[4]See discussion on the validity of text prompts in Step 4 of §2.3.

[5]We do not include actions in our concept list because actions are usually restricted to a small subset of objects (e.g., most objects cannot "cut", "dance" or "fly".).

[6]If there aren't enough existing objects for binding the spatial concepts, we request GPT-4o to add objects that naturally fit into the scene.

4

Step 3: Prompt Generation. Given the $k + 1$ statements and the binding structure represented in JSON format, GPT-4o is asked to make up a human-annotated description of a hypothetical image that matches the statements and the JSON object. GPT-4o is instructed to avoid introducing unnecessary objects or descriptions, as detailed in the prompting template in Appendix B.

Step 4: Prompt Validation. Before we feed the text prompts to T2I models, we have a prompt rejection mechanism to validate the text prompts with GPT-4o to rule out text prompts with hard conflict between visual concepts. Note that we do not simply remove unrealistic prompts (e.g., a horse with glass texture, as shown in Fig. 2), as they can be utilized to test the creativity of T2I models. As another example, this step rejects text prompts requesting a triangle-shaped person but keeps text prompts requesting a square-shaped cloud[7]. GPT-4o is asked to provide an explanation if it considers the text prompt invalid.

2.4 Concept Evaluation

We evaluate the generated images from T2I models by utilizing the visual question-answering capability of GPT-4o. Specifically, for each statement used in text prompt generation, we first ask GPT-4o to generate the corresponding yes or no question based on both the statement and the text prompt, and then send the question with the generated image to GPT-4o in a new conversation and record its answer ("Yes" or "No"). We award one point for each correctly illustrated statement, so the maximum possible points is $k + 1$.

Note naively asking GPT-4o or other vision language models (VLMs) whether the generated image matches the text prompt *does not work well* from our preliminary experiments, especially when k is large and the text prompts are complicated. Decomposing the text prompt is often used as an alternative for evaluating images generated from text prompts [8, 18]. However, previous decomposing methods may generate nonsensical questions when handling complex prompts [28], and thus harm their accuracy. Since the text prompts used in CONCEPTMIX are generated from given concepts, we have effectively decomposed the text prompt correctly. Although there might be additional information injected during our text prompt generation pipeline, we ensure the information injection is minimal and natural at each step. Our approach provides a reliable and precise method for evaluating the generated images based on the decomposed concepts from the original text prompt.

3 Experiments

In this section, we present a systematic evaluation of eight T2I models on CONCEPTMIX, with the experimental setup detailed in §3.1. We begin by analyzing the performance of individual concept categories ($k = 1$, see §3.2) to assess how well models handle specific concept categories in isolation. Next, we evaluate the models' performance when combining multiple concept categories ($k > 1$, see §3.3), and compare CONCEPTMIX with other existing evaluation pipelines (§3.4). Finally, we explore whether common training datasets are sufficient for effective compositional generation (§3.5).

3.1 Experimental Setup

Evaluated models. We evaluate eight state-of-the-art T2I models: SD v1.4 [35], DeepFloyd IF XL v1, SD v2.1, SDXL Base [33], SDXL Turbo [39], Playground v2.5 [26], PixArt alpha [7] and DALL·E 3 [2]. We provide the details of generation configuration and compute details for our evaluation in Appendix C.

Prompt Generation Details. We randomly generate text prompts from CONCEPTMIX, as detailed in §2.3, and request models for generations. Each prompt includes at least one object along with k additional visual concept categories. Unless specified otherwise, we randomly assign concepts from each category. We evaluate with $k \in \{1, 2, 3, 4, 5, 6, 7\}$, and for each k, we generate 300 text prompts to capture the variability and performance across different models.

---

[7]Because clouds can naturally have various abstract shapes, but a triangle-shaped person conflicts with the perceptual constraints on human form.

5

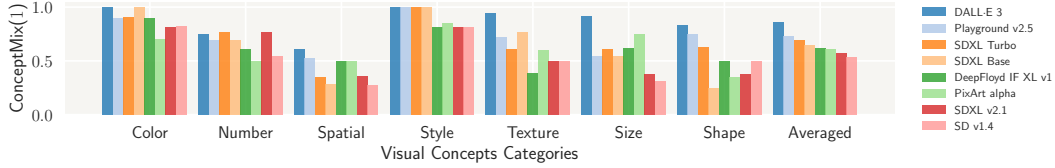



Figure 3: **Performance Across Concept Categories.** We evaluate the performance of T2I models across different concept categories. Color and style are easier, with all models achieving high scores. Performance is lower for generating specific numbers of objects and spatial relationships, with varying results for texture and size. Overall, DALL·E 3 outperforms others in all categories.

Concept Evaluation Details. Given a fixed k, we use GPT-4o, as described in §2.4, to grade each image and determine the number of points awarded out of $k + 1$, with each point representing a required concept. We consider two grading metrics: 1) **Full mark score**, which measures the proportion of generated images where the image correctly satisfies *all* $k + 1$ required concepts, and 2) **Concept fraction score**, which measures the average proportion of visual concepts satisfied by the generated images. Unless otherwise specified, the term 'performance' refers to full mark score. For each model and each k, we report the full mark score (Tab. 3) and concept fraction score (Appendix D), aggregated over 300 sampled prompts, and provide the 95% confidence interval for each score.

3.2 Performance on Individual Concept Categories ($k = 1$)

We begin by analyzing the performance of the models on the case $k = 1$ with each concept category, i.e., the ability to generate images of a random object and a concept within the selected category. This is the simplest form of compositional image generation. Our findings are listed as follows.

Color and style are easiest while spatial, size, and shape are challenging. Fig. 3 shows each model's performance across categories. A notable trend is that color and style are easier categories than others. For instance, DALL·E 3 excels in color and style, achieving perfect scores, and performs well in texture as well. However, it scores considerably lower in number and spatial categories, achieving only 0.75 and 0.61, respectively. Such findings highlight the limitations of using pixel-level similarity scores for evaluation. While these scores effectively capture style and color accuracy, they struggle to accurately reflect spatial, shape, and size. Consequently, models that perform well on these scores might still fall short in accurately generating spatial, shape, and size information.

Varying performance of concepts within the same category. Fig. 4 shows the performance of Playground v2.5 across different concepts within the easiest (color) and most challenging (spatial) categories identified earlier. The performance on different concepts varies significantly. In the color category, 'red' and 'green' score higher than 'brown' and 'black'. Similarly, for spatial concepts, 'in front of' and 'right' outperform 'left' and 'bottom'. Similar variations are observed in other categories with other models,

Figure 4: **Individual Concept Performance.** CONCEPTMIX scores for Playground v2.5 with $k = 1$ for colors (left) and spatial (right) concepts show performance varies within each category. More details on other categories are in Appendix D.

suggesting the existence of disparities in generation performance even within the same visual concept category. Based on the observation, we split each concept category into an easy subset and a hard subset. We then create two variants of CONCEPTMIX: one using the easy concepts and the other using hard concepts, see Appendix B for more details.

3.3 Performance of Compositional Generation ($k > 1$)

Models performance degrades when k increases. Now we examine model performance when combining multiple concept categories ($k > 1$) on our CONCEPTMIX benchmark. As shown in Tab. 3, DALL·E 3 consistently outperforms other models across all k difficulty levels and can handle complex compositional tasks more effectively. As k increases, all models show a significant drop in performance. Among all, the performance of SD v1.4 decreases the fastest as k increases, as we can see its performance approaching zero when $k = 3$. Other models also experience performance drops but at different rates. The models can be roughly ranked by their position in the table, with DALL·E 3 being the best, and SD v1.4 being the worst. SDXL Turbo, PixArt alpha, SDXL Base, DeepFloyd IF

6

k=1 The image shows two pigs.

k=2 A black giraffe drawn in ink.

k=3 A square-shaped teddy bear in a cartoon style is positioned on the left side of a ball.

k=4 A heart-shaped yellow sushi with a metallic texture is depicted in a photorealistic style.

k=5 The image shows three brown, circle-shaped robots with a fluffy texture. These robots are positioned behind a box.

k=6 Three tiny, rectangle-shaped pizzas with a metallic texture are positioned in front of a box. The scene is in an oil painting style

k=7 The image shows two huge red apples positioned on the right side of three blue necklaces.

SD v1.4 · SD v2.1 · PixArt Alpha · Playground v2.5 · DALL·E 3

Figure 5: **Qualitative performance** of different T2I models (DALL·E 3, PixArt alpha, Playground v2.5, SD v2.1, SD v1.4) across varying levels of compositional complexity ($k = 1...7$). As prompts become more complex, image quality degrade. DALL·E 3 performs best, while SD v1.4 performs worst.

Table 3: **Performance of Eight T2I Models on CONCEPTMIX.** We vary difficulty levels k from 1 to 7 and report the full mark scores, which represent the proportion of generated images that correctly satisfy all $k + 1$ required visual concepts. As k increases, all models' performance decreases, but at varying rates.

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
SD v1.4 [35]	$0.52_{\pm0.06}$	$0.23_{\pm0.05}$	$0.08_{\pm0.04}$	$0.03_{\pm0.03}$	$0.01_{\pm0.02}$	$0.00_{\pm0.01}$	$0.00_{\pm0.01}$
SD v2.1 [33]	$0.52_{\pm0.06}$	$0.29_{\pm0.05}$	$0.14_{\pm0.04}$	$0.06_{\pm0.03}$	$0.03_{\pm0.03}$	$0.01_{\pm0.02}$	$0.00_{\pm0.01}$
SDXL Turbo [39]	$0.64_{\pm0.06}$	$0.35_{\pm0.06}$	$0.18_{\pm0.05}$	$0.09_{\pm0.04}$	$0.03_{\pm0.03}$	$0.02_{\pm0.02}$	$0.01_{\pm0.02}$
PixArt alpha [7]	$0.66_{\pm0.06}$	$0.37_{\pm0.06}$	$0.17_{\pm0.05}$	$0.09_{\pm0.04}$	$0.05_{\pm0.03}$	$0.01_{\pm0.02}$	$0.01_{\pm0.02}$
SDXL Base [33]	$0.69_{\pm0.06}$	$0.43_{\pm0.06}$	$0.18_{\pm0.05}$	$0.09_{\pm0.04}$	$0.05_{\pm0.03}$	$0.01_{\pm0.02}$	$0.00_{\pm0.01}$
DeepFloyd IF XL v1 [43]	$0.68_{\pm0.06}$	$0.38_{\pm0.06}$	$0.21_{\pm0.05}$	$0.09_{\pm0.04}$	$0.05_{\pm0.03}$	$0.02_{\pm0.02}$	$0.01_{\pm0.02}$
Playground v2.5 [26]	$0.70_{\pm0.06}$	$0.46_{\pm0.06}$	$0.22_{\pm0.05}$	$0.10_{\pm0.04}$	$0.07_{\pm0.04}$	$0.02_{\pm0.02}$	$0.00_{\pm0.01}$
DALL·E 3 [2]	$0.83_{\pm0.05}$	$0.61_{\pm0.06}$	$0.50_{\pm0.06}$	$0.27_{\pm0.05}$	$0.17_{\pm0.05}$	$0.11_{\pm0.04}$	$0.08_{\pm0.04}$

XL v1, and Playground v2.5 have relatively close performance, with SDXL Base performing better at $k = 2$, DeepFloyd IF XL v1 and Playground v2.5 performing better at $k = 3$. We provide qualitative examples in Fig. 5 and we report the concept fraction score in Appendix D.

Easy and hard variants of CONCEPTMIX. We create two variants of CONCEPTMIX based on §3.2: one only uses the easy subsets of all categories, and the other uses the hard subsets. In Fig. 6, we plot the performance of three models on the two variants, as well as the standard CONCEPTMIX. With both variants, we again observe the degradation of model performance when k increases. Furthermore, the model ranking remains consistent, indicating the robustness of CONCEPTMIX. Although the easy and hard subsets are selected based on Playground v2.5 performance on these concepts with $k = 1$, models always achieve higher scores on the easy variant compared to the hard variant.

Figure 6: CONCEPTMIX(k) drops significantly as k increases, with DALL·E 3 consistently outperforming others. Shaded areas indicate the score range from easier to harder visual concepts for each k.

Legend: DALL-E 3; DALL-E 3 (Easy to Hard); Playground v2.5; Playground v2.5 (Easy to Hard); SD v1.4; SD v1.4 (Easy to Hard)

3.4 CONCEPTMIX has stronger discriminative power than other evaluation pipelines

We compare CONCEPTMIX with the prior compositional generation benchmark, T2I-CompBench [19], which uses a fixed template to combine at most five visual concept categories within a single prompt (see Tab. 1). While T2I-CompBench incorporates several evaluation metrics, its limited concept and prompt diversity often leads to closely clustered scores for different models, making it

7

Figure 7: **CONCEPTMIX Shows Stronger Discriminative Power.** We compare five models using 3-in-1 and GPT4v scores (global prompt-level) from T2I-CompBench [19], and CONCEPTMIX with varying difficulty levels (k). Unlike T2I-CompBench, which shows similar scores across models, CONCEPTMIX effectively differentiates model performance, with gaps widening as k increases.

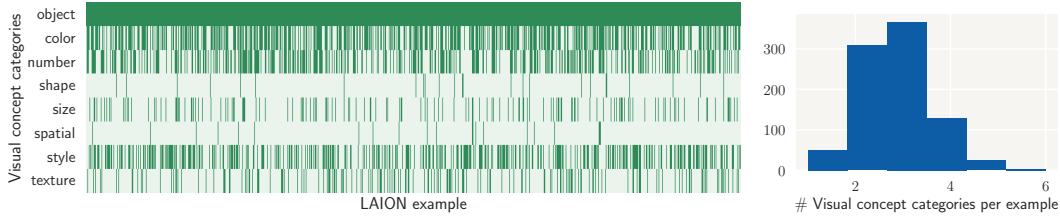



Figure 8: **Concept Diversity in LAION-5B Dataset.** Left: Heatmap of sampled captions shows colors and styles are most frequent; shapes and spatial relationships are least. Right: Most examples include 2-3 concepts.

challenging to differentiate their performance (see Fig. 7). This lack of differentiation also hinders the identification of model limitations. In contrast, CONCEPTMIX includes a wider range of concept categories and prompting variations (see Appendix B), and offers a more **precise** and **discriminative** grading approach (see Fig. 7), especially as k increases.

3.5 Tracing the poor performance of models back to lack of diversity in training data

To further investigate the relatively poor performance of models on CONCEPTMIX, we explore whether the complexity of visual concepts in the training data might be a contributing factor. We randomly sample 1000 image captions from LAION-5B [40], a widely used dataset for training T2I models. For each caption, we use GPT-4o to identify the presence of eight visual concepts categories[8]: object, color, number, shape, size, spatial, style, and texture. We filter out captions that did not contain objects (leaving 882 out of 1000) and plot the frequency of each concept in Fig. 8.

Disparate concept representation in LAION-5B. Our analysis reveals a significant disparity in the presence of different visual concepts within the LAION-5B dataset. While most captions included color (476) and style (269), only a small number contained shape (24) and spatial (20) concepts. This uneven distribution aligns with the individual visual concept performance observed in Section 3.2, suggesting that a model's proficiency in a particular visual concept might be directly influenced by the frequency of its representation in the training data.

Limited exposure to complex concept combinations in LAION-5B. Furthermore, we find that each example from the sampled LAION-5B collection, on average, contains only 2.75 ± 0.90 concept categories, with a maximum of six concepts per example. This limited exposure to complex combinations of visual concepts in the training data likely contributes to the observed difficulty models face when dealing with $k \geq 3$ (see Tab. 3).

4 Related Work

Compositional Generalization. Compositionality is key to generalizing existing knowledge to new tasks and therefore has attracted significant attention in machine learning. In CV, studies have explored compositional generalization in disentangled representation learning [16, 11, 46], visual relations [29], as well as concept compositions [32]. Other works focus on compositional models for image generation [10], and planning for unseen tasks at inference time [9]. In NLP, compositional generalization has also been studied extensively [13, 24, 4, 20, 21, 30]. SKILL-MIX [47], a more recent evaluation on LLMs, presented a more general approach to evaluate compositional generalization.

---

[8]Instructions for GPT-4o are provided in Appendix C.

8

SKILL-MIX asks LLMs to produce novel pieces of text from random combinations of k skills, which can be made more difficult by simply increasing the value of k. CONCEPTMIX is partly inspired by SKILL-MIX, but requires a more complicated design in creating text prompts and effective grading.

T2I models and compositional T2I benchmarks. T2I models [35, 2, 3, 5, 33, 43, 26] generate images given text prompts. Traditionally, their performance is evaluated based on alignment with reference (image, caption) pairs. This involves querying the T2I model with the reference caption and assessing the consistency between the generated image and the reference image. Common benchmarks include TIFA160 [18], Pick-a-Pic [22], DrawBench [37], and COCO-T2I [27]. When reference images are not provided, benchmarks with prompt templates are used for a more comprehensive measure of compositional capabilities [12, 5, 1, 19, 25]. Among them, the closest to ours is T2I-CompBench [19], which samples complex prompts to evaluate T2I models. However, as noted in Tab. 1, T2I-CompBench limits prompts to 5 concepts, while CONCEPTMIX uses up to 8 (i.e., $k = 7$).

Evaluation metrics for generation. Most previous benchmarks use similarity metrics like Inception Score [38, IS], Fréchet Inception Distance [15, FID], and Learned Perceptual Image Patch Similarity [48, LPIPS] to quantify generation quality. These metrics, relying on pre-trained networks, primarily capture pixel-level similarity and often fail to fully capture semantic-level alignment. To address these limitations, recent methods [41, 45, 36] have adopted metrics like CLIPScore [34, 14], which measure cosine similarity between embedded image and text representations, and visual question answering pipelines [23, 49, 28] to better capture text-image alignment. Our evaluation also adopt the visual question answering pipeline for text-image consistency checking, but with a more careful design of asking appropriate questions to verify the generation quality of each visual concept thanks to our prompt generation pipeline.

5 Discussion

Limitations. One potential limitation of our CONCEPTMIX benchmark is the possible misalignment between autograding and human grading. While our grading method shows great improvement and aligns with human preference (Appendix A) compared to previous metrics, it may not always capture the details that a human grader would and might miss or misinterpret some questions. This discrepancy could lead to differences in scores, particularly in cases where the generated images are ambiguous. Therefore, while our grading engine offers consistent and scalable evaluation, outperforming previous approaches, it still cannot fully replicate human judgment.

Negative Impacts. T2I generation via models trained on web-scale data carries inherent risks, such as privacy and copyright violations, or the perpetuation of social bias. Although our work focuses on the *evaluation* of the generative models, with the goal of reducing errors in generation, the downside is that CONCEPTMIX may also provide further legitimacy to generative models despite their underlying ethical concerns.

6 Conclusion

Compositional capabilities are critical for T2I generation. We gave evidence that existing evaluations of compositionality, which generate prompts automatically with fixed templates, actually result in prompts with low diversity and discriminative power. We propose CONCEPTMIX, a scalable and customizable benchmark for evaluating the compositional capabilities of T2I models, including prompts from 8 visual concept categories. Our approach uses a powerful LLM in two ways to address the limitations of existing benchmarks. The first is in generating suitable prompts given a random set of visual concepts. The second is to enable automated grading of the generated image by providing a list of questions that can be used with a VLM (GPT-4o in our case) to check the correctness of the generated images. CONCEPTMIX allows generating a wide variety of prompts — the total number of possible prompts is larger than the size of popular training datasets. We find that CONCEPTMIX effectively differentiates between models, offering a more granular understanding of the strengths and weaknesses of generation models compared to traditional benchmarks.

9

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. 2107048. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. DY and SA are supported by NSF and ONR. YH is supported by the Wallace Memorial Fellowship.

References

[1] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023.

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

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[4] Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, 2020.

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[7] 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.

[8] 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. *arXiv preprint arXiv:2310.18235*, 2023.

[9] Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need, 2024.

[10] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation and inference with energy based models, 2020.

[11] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2525–2534. PMLR, 2019.

[12] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.

[13] Catherine Finegan-Dollak, Jonathan K Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. Improving text-to-sql evaluation methodology. *arXiv preprint arXiv:1806.09029*, 2018.

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[15] 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. *Advances in neural information processing systems*, 30, 2017.

10

[16] Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*, 2017.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[18] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.

[19] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[20] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: how do neural networks generalise?, 2020.

[21] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data, 2020.

[22] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

[23] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023.

[24] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

[25] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36, 2024.

[26] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

[27] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[28] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024.

[29] Nan Liu, Shuang Li, Yilun Du, Joshua B. Tenenbaum, and Antonio Torralba. Learning to compose visual relations, 2021.

[30] Qian Liu, Shengnan An, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Nanning Zheng, and Dongmei Zhang. Compositional generalization by learning analytical expressions. *Advances in Neural Information Processing Systems*, 33:11416–11427, 2020.

[31] OpenAI. Hello gpt-4o, 2024.

[32] Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14554–14562, 2024.

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

11

[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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[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. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[39] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

[40] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[41] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[42] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[43] StabilityAI. DeepFloyd IF. https://github.com/deep-floyd/IF, 2023.

[44] Zhen Wang, Yuelei Li, Jia Wan, and Nuno Vasconcelos. Diffusion-based data augmentation for object counting problems. *arXiv preprint arXiv:2401.13992*, 2024.

[45] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[46] Zhenlin Xu, Marc Niethammer, and Colin A Raffel. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. *Advances in Neural Information Processing Systems*, 35:25074–25087, 2022.

[47] Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-mix: A flexible and expandable family of evaluations for ai models. *arXiv preprint arXiv:2310.17567*, 2023.

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[49] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

12

Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We provide the paper's contributions and scope and emphasizing the challenges of our benchmark in the abstract and introduction.

   (b) Did you describe the limitations of your work? [Yes] We provide the limitation discussion in Sec. 5.

   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We provide this discussion in Sec. 5.

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [N/A]

   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide the code, data and instruction needed in the supplemental material.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We add the experiment details in the appendix Sec. C.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report the error bars in Tab. 3.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provide those details in Appendix Sec. C.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite all of the works that we used in our work.

   (b) Did you mention the license of the assets? [Yes] We mention that the LAION dataset we analyzed has MIT License

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Yes, we provide our dataset and coadebase in the supplemental material.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] All our data are generated and they are not from other people.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Our data are synthetic data and do not contain personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] We include them in Appendix A.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

13

Appendices

A Human Evaluation   15

   A.1 Setup   15

   A.2 Human Evaluation Instructions   15

   A.3 Results   16

   A.4 Feedback from human annotators   22

B Benchmark Details   23

   B.1 Configuration Details   23

   B.2 Prompt Generation   23

   B.3 Question Generation   25

C Experimental Details   26

   C.1 Compute Resource   26

   C.2 Generation Configurations   26

   C.3 Experimental details for §3.5   26

D Additional Experimental Results   27

E Common Failure Cases   28

   E.1 Numbers   28

   E.2 Shapes   30

   E.3 Sizes   32

   E.4 Textures   34

   E.5 Spatial Relationship   36

   E.6 Styles   38

   E.7 Colors   40

14

518 **We release our code here:** `https://github.com/princetonvisualai/ConceptMix`.

A Human Evaluation

A.1 Setup

521 To evaluate the performance of our automatic grading with GPT-4o, we conduct human evaluation
522 experiments. Each pair of generated results was evaluated by nine participants, including both experts
523 in the field and individuals without specific background knowledge, two of the participants are authors
524 of this paper. We conduct human evaluation for 14 sets: $k = 3$ across all eight evaluated models and
525 $k = 1, ..., 7$ for DALL·E 3. Each set includes 25 pairs of text prompts and generated images, resulting
526 in 350 pairs in total.

A.2 Human Evaluation Instructions

528 Here are the instructions for participants in the human evaluation:

Human Evaluation Instructions

Your task is to evaluate the alignment between the image and the text description. Follow the steps
outlined below:

Step 1: Judge the Alignment. First, determine whether the image aligns with the description provided
in the prompt. If the image aligns with the description, proceed to Step 2. If the image does not align
with the description, your answer should be 0 (no).

Step 2: Double-Check the Answers. If you determined that the image aligns with the description
in Step 1, then verify if all the specific questions listed are correctly answered with "yes" or "no". If
all answers to the questions are "yes", then your final answer should be 1 (yes). If any answer to the
questions is "no", then your final answer should be 0 (no).

Example:
Step 1: Judge the Alignment
Prompt: A photorealistic image shows a rectangle-shaped smartphone positioned in front of a table,
closer to the observer. The smartphone is clearly distinguishable from the table behind it.

If you answered 1 (yes): then do Step 2, otherwise directly answer 0 (no).

Step 2: Double-Check the Answers; check whether all answers are correct, if yes → 1, if any answer is
incorrect → 0.
Question #1: Does the image contain a smartphone?
Question #2: Is the style of the image photorealism?
Question #3: Is the smartphone rectangle-shaped?
Question #4: Is the smartphone positioned in front of the table, closer to the observer?

---

530 In addition to the instructions and example above, we also offer general guidance for visual concepts
531 that may be subjective in judgment. Specifically,

532 **Size** For "tiny" and "huge", judge whether the object is tiny or huge compared to its normal size in
533   reality, which can be inferred based on the size of other objects (assuming the other objects
534   have normal sizes).

535 **Style** We define all the art styles in the rubric and provide reference images.

Back to Table of Contents   15   Back to the First Page

536 **A.3   Results**

537 **GPT-4o grader in general shows high consistency with human annotators.** Fig. 9 presents the
538 consistency scores among human annotators and between human annotators and GPT-4o. Consistency
539 score is defined as the ratio of two scorers giving the same score for a (prompt, image) pair among all
540 of the (prompt, image) pairs. As illustrated, the average consistency score between human annotators
541 for this task is 0.75, showing the relative subjectivity and challenge of the evaluation. In contrast, the
542 consistency score between the human majority vote and GPT-4o is 0.82, indicating that GPT-4o is
543 more aligned with the consensus of human annotators than the human annotators are with each other
544 on this task.

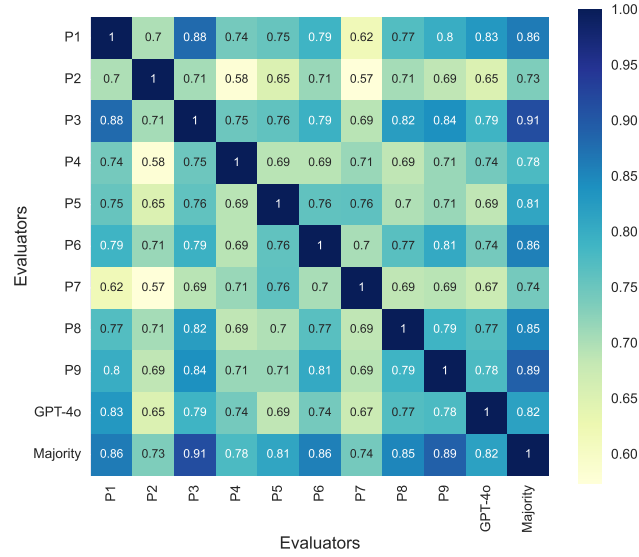



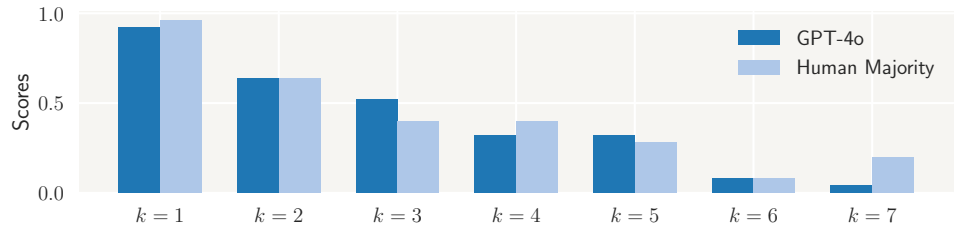
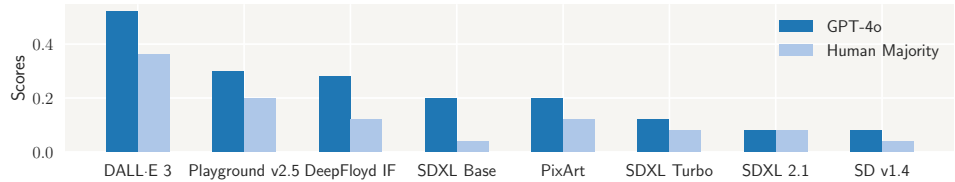
Figure 9: **Pairwise Consistency Heatmap.** The heatmap shows the consistency between different human
evaluators (**P1** to **P9**) as well as a majority vote (**Majority**) and GPT-4o (**GPT-4o**) across all k for DALL·E 3.
Each cell represents the consistency score, with darker shades showing higher agreement between evaluators.
The average human-to-human consistency is 0.75, which reveals that human evaluations also vary a lot compared
to automated evaluation methods.

(a) GPT-4o and human scores for DALL·E 3 model generations on CONCEPTMIX with different k

(b) GPT-4o and human scores for generations on CONCEPTMIX with $k = 3$ across different models

Figure 10: **Our Scores vs. Human Scores** on CONCEPTMIX with (a) different k values for the DALL·E 3
model, and (b) $k = 3$ for different models.

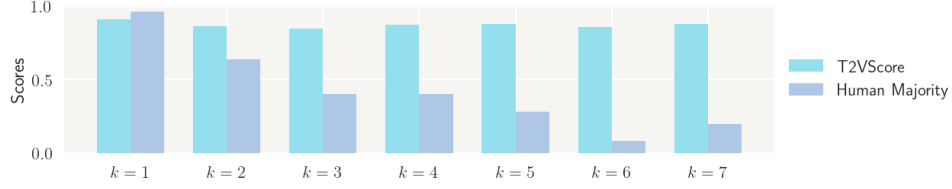
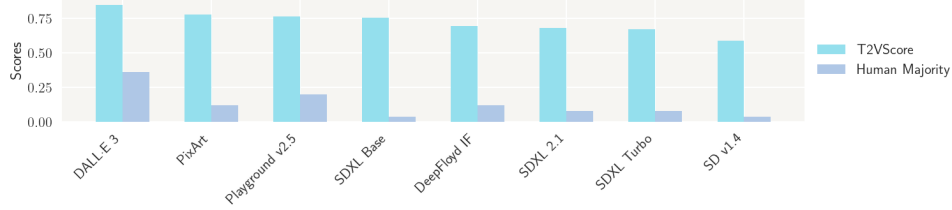
Back to Table of Contents   16   Back to the First Page

(a) T2VScore [28] and human scores for DALL·E 3 model generations on CONCEPTMIX with different k

(b) T2VScore [28] and human scores for generations on CONCEPTMIX with $k = 3$ across different models

Figure 11: **T2VScore [28] vs. Human Scores** on CONCEPTMIX with (a) different k values for the DALL·E 3 model, and (b) $k = 3$ for different models.

In Fig. 10, we compare the full mark scores by GPT-4o and human scores over different settings. Human scores are the average of the human majority votes across 25 pairs. From Fig. 10a, we observe that GPT-4o is close to human scores, except for $k = 7$, the human annotators give much higher scores than the GPT-4o. It may be caused by human oversight when the complexity of text prompts increases. Despite this, the overall trend of human scores shows a decline as k increases, matching the trend of GPT-4o scores. In Fig. 10b, we sort the models by their GPT-4o scores. We observe that the human ranking is similar to GPT-4o ranking except SDXL Base. Additionally, human annotators consistently give lower scores than GPT-4o, which is likely because human annotators are more familiar with these text prompts as they are identical for all models.

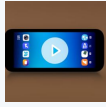
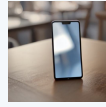
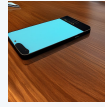
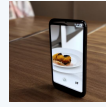
Compare with Prior Grading Approach. We further conduct experiments with previous state-of-the-art grading approach [28] and compare them with human preferences. As shown in Fig. 10 and Fig. 11, our grading method aligns better with human preferences, for example, in Fig. 10a, as k grows, both our grading results and human majority vote results generally decrease. However, this trend is not observed in Fig. 11a, and T2VScore barely changes when k grows. Additionally, in Fig. 11b, where we sorted the models by their T2VScore performance, we observe that T2VScores are again similar for many models, and human scores do not correlate with it well. This shows that our grading approach can differentiate between various generation models and better reflect human preferences. Our method stands out by accounting for different difficulty levels and providing a detailed understanding of model performance.

Qualitative Analysis. During the evaluation, we noticed several instances where human evaluators disagreed among themselves or with the GPT-4o grading method. In some cases, GPT-4o tends to be stricter in its grading. For instance, an image slightly deviating from the prompt's specifics might receive a lower score from GPT-4o, while human evaluators might overlook minor discrepancies and incorrectly grade it higher. Here we show some examples:

Back to Table of Contents   17   Back to the First Page

Human-GPT-4o Disagreement Example 1 (k=3)

Prompt: A photorealistic image shows a rectangle-shaped smartphone positioned in front of a table, closer to the observer. The smartphone is clearly distinguishable from the table behind it.

| DALL·E 3 | Playground v2.5 | DeepFloyd IF XL v1 | SDXL Base | PixArt Alpha | SDXL Turbo | SD v2.1 | SD v1.4 |

Grading results:

```
Human (9 participants):
    P1: 1 0 0 0 0 0 0 1
    P2: 1 1 1 1 1 0 1 1
    P3: 0 0 0 0 0 0 0 1
    P4: 0 1 0 1 1 1 1 0
    P5: 1 1 1 1 1 1 1 1
    P6: 0 1 1 1 1 1 1 0
    P7: 1 0 0 0 0 0 0 1
    P8: 0 0 0 0 0 0 0 0
    P9: 0 0 0 0 0 0 0 1
GPT-4o: 0 0 0 1 1 1 1 1
```

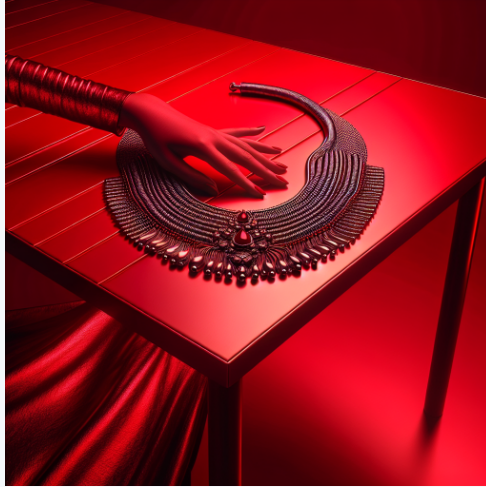
Automatic grading questions:

```
Does the image contain a smartphone?
Is the style of the image photorealism?
Is the smartphone rectangle-shaped?
Is the smartphone positioned in front of the table, closer to the observer?
```

569

Back to Table of Contents   18   Back to the First Page

Human-GPT-4o Disagreement Example 2 (DALL·E 3, k=4)

Prompt: The image shows a red table with a red metallic-textured necklace placed on its surface.

Grading results:

```
Human (9 participants):
1 1 1 1 0 1 1 0 1
GPT-4o: 0
```

GPT-4o grading details:

```
Does the image contain a table?        1
Does the image contain a necklace?      1
Is the color of the necklace red?       0
Is the color of the table red?          1
Does the necklace have a metallic texture? 1
```

570

Back to Table of Contents   19   Back to the First Page

Human-GPT-4o Disagreement Example 3 (DALL·E 3, k=5)

Prompt: A tiny elephant is positioned to the left of a tiny white broccoli.

Grading results:

```
Human (9 participants):
1 0 0 1 0 1 1 1 0
GPT-4o: 0
```

GPT-4o grading details:

```
Does the image contain an elephant?                      1
Does the image contain a broccoli?                       1
Is the elephant tiny?                                    1
Is the color of the broccoli white?                      0
Is the broccoli tiny?                                    0
Is the elephant positioned on the left side of the broccoli? 1
```

571

Back to Table of Contents   20   Back to the First Page

Human-GPT-4o Disagreement Example 4 (DALL·E 3, k=6)

Prompt: The image shows a blue robot with a glass texture positioned to the right of a tiny rose. The style of the image is photorealism.

Grading results:

```
Human (9 participants):
1 0 0 0 1 1 0 1 0
GPT-4o: 0
```

GPT-4o grading details:

```
Does the image contain a robot?                        1
Does the image contain a rose?                         1
Is the size of the rose tiny?                          0
Is the color of the robot blue?                        1
Is the style of the image photorealism?                0
Does the robot have a glass texture?                   1
Is the robot positioned on the right side of the rose? 0
```

572

Back to Table of Contents   21   Back to the First Page

Human-GPT-4o Disagreement Example 5 (DALL·E 3, k=7)

Prompt: On a large plate, there is a heart-shaped piece of sushi. Next to it, there is a fork with a glass texture. A tiny butterfly is perched on the edge of the plate. Nearby, a cactus with a fluffy texture is also present.

Grading results:

```
Human (9 participants):
1 0 0 1 1 1 0 0 0
GPT-4o: 0
```

GPT-4o grading details:

```
Does the image contain a fork?         1
Does the image contain a butterfly?    1
Does the image contain sushi?          1
Does the image contain a cactus?       1
Is the sushi heart-shaped?             1
Does the fork have a glass texture?    0
Is the butterfly tiny?                 0
Does the cactus have a fluffy texture? 0
```

These results highlight the challenges of achieving high inter-human rater reliability in subjective evaluations and show the strengths of our automatic grading method with GPT-4o.

A.4 Feedback from human annotators

We received feedback from human annotators and listed details below.

- There exists phrasing with ambiguity, e.g., in the first example of §A.3, whether it requires the phone to be closer than the front edge of the table, or it covers some part of the table?

- Feedback related to styles: some of the styles are too difficult for models (e.g., expressionism), and some of the styles are difficult to judge (e.g., impressionism); some concepts are hard to realize in certain styles (e.g., "fluffy" texture in "cubism").

- Additional information injected by GPT-4o in prompt generation pipeline: some text prompts contain the quantifier "a single object" even though the individual questions do not require that.

In general, most annotators find some images hard to grade and some questions hard to answer, which is aligned with relatively low consistency between annotators, observed from Fig. 9. All feedback provides useful insights for future updates of CONCEPTMIX and the development of similar benchmarks.

Back to Table of Contents   22   Back to the First Page

B Benchmark Details

B.1 Configuration Details

Below are the detailed concept values for each visual concept category in CONCEPTMIX:

Objects: apple, bee, broccoli, butterfly, cactus, car, carrot, cat, chair, chicken, corgi, cow, dirt road, doll, dog, duck, elephant, fork, giraffe, hammer, highway, hill, house, laptop, lion, man, necklace, novel, oak tree, orange, pig, pine tree, pizza, ring, robot, rose, screwdriver, sheep, skyscraper, smartphone, spider, spoon, sunflower, sushi, table, teddy bear, textbook, truck, woman, zebra

Colors: black, blue, brown, gray, green, orange, pink, purple, red, white, yellow

Numbers: 2, 3, 4

Shapes: circle, heart, rectangle, square, triangle

Sizes: huge, tiny

Textures: fluffy, glass, metallic

Spatial Relationship: above, behind, below, bottom, in front of, inside, left, outside, right, top

Styles: abstract, cartoon, cubism, expressionism, graffiti, impressionism, ink, manga, oil painting, photorealism, pixel art, pop art, sketch, surrealism, watercolor

Values in blue indicate easy splits, while values in orange denote hard splits of different concepts, as measured on Playground v2.5 with $k = 1$. We use these splits for experiments in §3.3. Note that we use all objects for both easy and hard splits to ensure a fair comparison.

B.2 Prompt Generation

We use GPT-4o (endpoint of May 13th, 2024), to help bind multiple concepts and generate prompts, as detailed in §3.3. For concept bind, we utilize the JSON format, and start with a JSON in the following structure:

Example of Initial JSON for concept binding

```
{"objects": [{"id": 1, "item": "teddy bear", "color": "green", "texture":
"glass", "number": "4"}, {"id": 2, "item": "laptop", "shape": "rectangle",
"size": "tiny"}], "style": "oil painting", "relation": [{"name": "behind",
"description": "{ObjectA} is behind {ObjectB}, meaning {ObjectA} is
positioned farther from the observer or camera than {ObjectB}", "ObjectA_id":
"?", "ObjectB_id": "?"}]}
```

We intentionally leave some question marks for spatial relationships, and ask GPT-4o to fill them and potentially add new objects if needed. The instruction given to GPT-4o is as follows:

Instructions given to GPT-4o for finalize JSON

I am trying to create an image containing exactly the following things in a JSON format:
[Initial JSON]
Could you check if there is "?" left in the JSON? If so, could you fill in the missing part? Make sure it makes sense when you fill the missing part. Do not fill in anything else unless it is indicated by "?". You may add additional objects, but only in the following two cases:
* It is needed to fill in any "?" (Note when you fill "?", you should use existing objects first. If you still choose to add an object, explain why the existing objects cannot fulfill the need.); or
* If there is an attribute specified in the JSON that contains relative information (e.g. "size") and there is no other object for reference. (The reason for adding an object for this case is because one cannot tell whether an object is huge without any other object in the image, but we are fine if there is no such attribute mentioned in the JSON. Note other existing objects in JSON can be used for reference, and the reference object does not need to be the same object. If you still choose to add an object, explain why the existing objects cannot fulfill the need.)
DO NOT add any object if none of the above situations is strictly satisfied, and DO NOT try to improve the image in other ways. If you choose to add an object, make sure it fits in the image naturally. Please only add the necessary objects, and the added objects should only have "id" and "item" specified, and should be appended to "objects".

Back to Table of Contents   23   Back to the First Page

616 After we obtain the final JSON, we use the following instructions to produce the text prompt:

617

> **Instructions given to GPT-4o for text prompt generation**
>
> Make up a human-annotated description of an image that describe the following properties (meaning you can infer these properties from the description):
> [description of properties]
> As a reference, I constructed a JSON containing all the information from the properties and some additional information that you should incorporate into your description:
> [final JSON]
> Describe the image in an objective and unbiased way. Keep the description clear and unambiguous, and synthesize the objects in a clever and clean way, so people can roughly picture the scene from your description. DO NOT introduce unnecessary objects and unnecessary descriptions of the objects beyond the given properties and JSON. If there is an interaction between two objects, make sure the two objects are distinguishable. Avoid any descriptions involving a group of objects, or an ambiguous number of objects like "at least one", "one or more", or "several". Do not add subjective judgments about the image, it should be as factual as possible. Do not use fluffy, poetic language, or any words beyond the elementary school level. Respond "WRONG" and explain if the properties have obvious issues or conflicts, or if it is hard to realize them in an image. Otherwise, respond only with the caption itself.

618 Here the property description of each selected concept category is generated using the template
619 provided in Tab. 4.

Table 4: Template to format selected concepts with their corresponding descriptions presented to GPT-4. Values in brackets [] represent chosen visual concepts from their respective categories.

Category	Description template
Objects	the image contains one or more [object name]
Colors	the color of [object name] is [color name]
Numbers	the number of [object name] is exactly [number]
Shapes	[object name] is [shape name] shaped
Sizes	[object name] has a [size value] size
Textures	[object name] has a [texture name] texture
Spatial, top	[Object A] is on top of [Object B], meaning [Object A] is positioned above or at the highest point of [Object B], touching each other
Spatial, botton	[Object A] is at the bottom of [Object B], meaning [Object A] is positioned below or at the lowest point of [Object B], touching each other
Spatial, above	[Object A] is above [Object B], meaning [Object A] is positioned higher than [Object B] without touching it
Spatial, below	[Object A] is below [Object B], meaning [Object A] is positioned lower than [Object B] without touching it
Spatial, left	[Object A] is positioned on the left side of [Object B]
Spatial, right	[Object A] is positioned on the right side of [Object B]
Spatial, behind	[Object A] is behind [Object B], meaning [Object A] is positioned farther from the observer or camera than [Object B]
Spatial, in front of	[Object A] is on top of [Object B], meaning [Object A] is positioned above or at the highest point of [Object B], touching each other
Spatial, inside	[Object A] is inside [Object B], meaning [Object A] is positioned within the boundaries or interior of [Object B]
Spatial, outside	[Object A] is outside of [Object B], meaning [Object A] is positioned beyond the boundaries or exterior of [Object B]
Styles	the style of the image is [style name]

620 After generating the prompts, we then prompt GPT-4o for validation (see §2.3), using the following
621 instruction:

> **Instructions given to GPT-4o for prompt validation**
>
> Could you read your caption again and verify if it makes sense in a very loose sense (e.g., a person cannot be triangle shaped, but a cloud can be square-shaped and a tree can be rectangle-shaped)? If yes, respond with the exact same caption. If not, respond with "WRONG" and explain why.

622

623 We then filter out prompts that receive a "WRONG" response.

624 **Prompt length.** We also provide the distribution of text prompt lengths for different values of k. The
625 length of the text prompt may indicate the complexity of the task, as longer prompts tend to involve
626 more concepts. The distribution of text prompt lengths for each k is shown in Fig. 12.

Back to Table of Contents   24   Back to the First Page

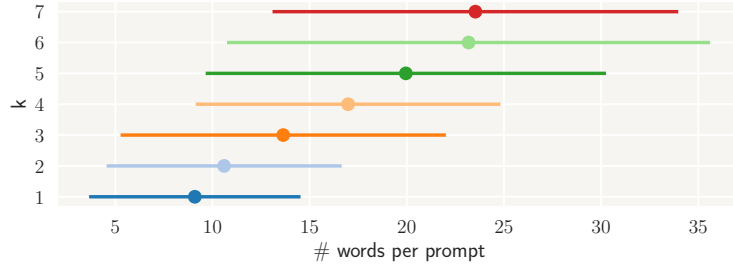



Figure 12: Distribution of prompt length in CONCEPTMIX: Larger values of k result in longer and potentially more complex prompts.

B.3   Question Generation

For each generated prompt, we also accompany it with a list of GPT-4o-generated questions, as detailed in §2.4, which are later used for grading. Specifically, we use the following instruction:

Instructions given to GPT-4o for question generation

A student just draw a picture based on your description. Can you help me verify whether the student did a good job? Specifically, I want to know if the image follows your description and also follows the properties I mentioned earlier. You should ask me one yes or no question for each property, and I will tell you if they are satisfied. For example, for properties like "the image contains one or more [object name]", the corresponding question should be "Does the image contain [object name]". Respond only the k questions, one for each property, in the same order of the properties, and each on a new line.

Back to Table of Contents   25   Back to the First Page

C Experimental Details

C.1 Compute Resource

All experiments are conducted on a single NVIDIA A6000 GPU card with 48GB memory. Tab. 5 provides statistics on the time cost for each image generation across all the evaluated models.

Table 5: Averaged time cost per generation for evaluated models using a single NVIDIA A6000 GPU card.

Model	Time cost (seconds) per generation
SD v1.4	2.17
SDXL Turbo	0.34
SD v2.1	3.99
SDXL Base	10.03
DeepFloyd IF XL v1	18.69
DALL·E 3	12.58
Playground v2.5	10.17
PixArt alpha	4.41

C.2 Generation Configurations

For all open-source models, we use their checkpoints from Hugging Face for generation, as listed in Tab. 6, with their default generation configurations. For DALL-E, we generate images via its API endpoint with the default settings[9].

Table 6: Summary of evaluated models with corresponding Hugging Face links and licenses.

Model	Hugging Face Link
SD v1.4	https://huggingface.co/CompVis/stable-diffusion-v1-4
SDXL Turbo	https://huggingface.co/stabilityai/sdxl-turbo
SD v2.1	https://huggingface.co/stabilityai/stable-diffusion-2-1
SDXL Base	https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
DeepFloyd IF XL v1	https://huggingface.co/DeepFloyd/IF-I-XL-v1.0
Playground v2.5	https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic/
PixArt alpha	https://huggingface.co/PixArt-alpha/PixArt-XL-2-1024-MS

(a) Models and their Hugging Face links

Model	License
SD v1.4	CreativeML OpenRAIL M license
SDXL Turbo	Stability AI Non-commercial Research Community License
SD v2.1	CreativeML Open RAIL++-M License
SDXL Base	CreativeML Open RAIL++-M License
DeepFloyd IF XL v1	DeepFloyd IF License Agreement
Playground v2.5	Playground v2.5 Community License
PixArt alpha	CreativeML Open RAIL++-M License

(b) Models and their licenses

C.3 Experimental details for §3.5

In §3.5, we analyze the concept diversity of LAION [40] (MIT License). We prompt GPT-4o to identify the number of visual concepts in each sampled caption from LAION:

> **Instructions given to GPT-4o for concept identification**
>
> Given a prompt, identify whether it includes any concept from the following visual concept categories: object, color, number, shape, size, spatial relationship, style, and texture. Directly return the included visual concept categories as your answer. If there is no detected visual concept categories, return an empty string.

---

[9] https://platform.openai.com/docs/api-reference/images/create

Back to Table of Contents   26   Back to the First Page

643
D   Additional Experimental Results

644
Following Fig. 4, we visualize all of the concept categories in Fig. 13.

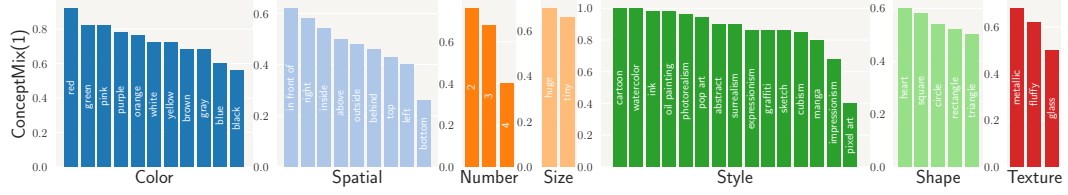



Figure 13: **Performance of concepts within the same category.**

645
Tab. 7 provides the concept fraction score of all evaluated models, showing a high correlation with the
646
full mark score reported in Tab. 3. Similar to Tab. 3, the concept fraction score drops when increasing
647
k, with DALL·E 3 being the best, and SD v1.4 being the worst. Note the drop in concept fraction
648
score not only indicates the difficulty level increase of the whole text prompts but also shows each
649
concept is harder to realize with more concepts described in the prompt.

Table 7: **Performance of T2I Models on our CONCEPTMIX benchmark.** Concept fraction score of seven
state-of-the-art T2I models with varying difficulty levels k from 1 to 7. As k increases, the performance of all
models decreases, but at different rates.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
SD v1.4 [35]	0.74 ± 0.03	0.61 ± 0.03	0.55 ± 0.03	0.50 ± 0.02	0.44 ± 0.02	0.41 ± 0.02	0.36 ± 0.02
SD v2.1 [33]	0.74 ± 0.03	0.68 ± 0.03	0.61 ± 0.03	0.54 ± 0.03	0.50 ± 0.03	0.48 ± 0.02	0.45 ± 0.02
SDXL Turbo [39]	0.81 ± 0.03	0.72 ± 0.03	0.65 ± 0.03	0.60 ± 0.03	0.57 ± 0.02	0.54 ± 0.02	0.49 ± 0.02
PixArt alpha [7]	0.82 ± 0.03	0.73 ± 0.03	0.67 ± 0.03	0.61 ± 0.03	0.56 ± 0.02	0.53 ± 0.02	0.49 ± 0.02
SDXL Base [33]	0.84 ± 0.03	0.76 ± 0.03	0.69 ± 0.02	0.63 ± 0.02	0.60 ± 0.02	0.57 ± 0.02	0.53 ± 0.02
DeepFloyd IF XL v1 [43]	0.84 ± 0.03	0.74 ± 0.03	0.66 ± 0.03	0.61 ± 0.02	0.59 ± 0.02	0.55 ± 0.02	0.51 ± 0.02
Playground v2.5 [26]	0.84 ± 0.03	0.77 ± 0.03	0.71 ± 0.02	0.64 ± 0.02	0.62 ± 0.02	0.58 ± 0.02	0.52 ± 0.02
DALL·E 3 [2]	0.92 ± 0.02	0.85 ± 0.02	0.83 ± 0.02	0.76 ± 0.02	0.75 ± 0.02	0.72 ± 0.02	0.71 ± 0.02

Back to Table of Contents   27   Back to the First Page

E   Common Failure Cases

In this section, we analyze frequent failure cases faced by T2I models, and we provide the visualizations of two failure cases across all visual concept categories.

E.1   Numbers

<div style="border:1px solid #1f77b4">

Numbers Failure Case (Example 1, Playground v2.5)

Prompt: The image shows four elephants and one zebra standing on a grassy plain.

Prompt Generation:

```
{
    "num_skills": 2,
    "categories": [
        "object", "object", "number"
    ],
    "skill": [
        "elephant", "zebra", "4"
    ]
}
```

Grading Results:

```
{
"questions": [
    "Does the image contain elephants?  ",
    "Does the image contain zebras?  ",
    "Does the image contain exactly 4 elephants?"
    ],
"scores": [
    1,
    0,
    0
    ]
}
```

</div>

Back to Table of Contents   28   Back to the First Page

Numbers Failure Case (Example 2, DALL·E 3)

Prompt: In a pop art style image, there are two huge glass-textured carrots. In front of the carrots, there are three tiny giraffes. Additionally, there is an apple included in the scene.

Prompt Generation:

```
{
    "num_skills": 7,
    "categories": [
        "object", "object", "number", "size", "number", "texture",
        "style", "size"
    ],
    "skill": [
        "carrot", "giraffe", "3", "tiny", "2", "glass", "pop art", "huge"
    ]
}
```

Grading Results:

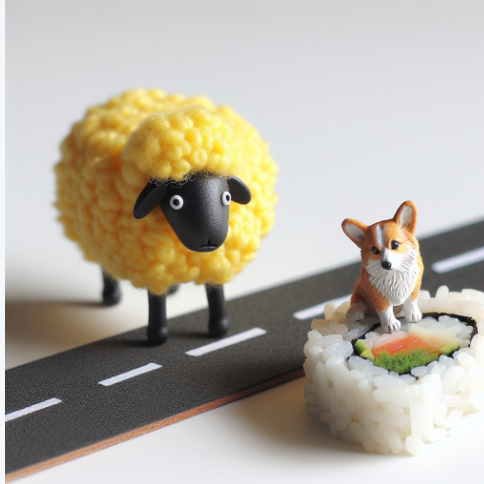
```
{
    "questions": [
        "Does the image contain one or more carrots?  ",
        "Does the image contain one or more giraffes?  ",
        "Does the image contain exactly 3 giraffes?  ",
        "Are the giraffes tiny in size?  ",
        "Does the image contain exactly 2 carrots?  ",
        "Do the carrots have a glass texture?  ",
        "Is the style of the image pop art?  ",
        "Are the carrots huge in size?"
    ],
    "scores": [
        1,
        1,
        0,
        1,
        1,
        0,
        0,
        1
    ]
}
```

655

Back to Table of Contents   29   Back to the First Page

656 ## E.2 Shapes

Shapes Failure Case (Example 1, DALL·E 3)

Prompt: A tiny yellow sheep stands on a heart-shaped highway. Nearby, a small corgi sits next to a piece of sushi.

Prompt Generation:

```
{
    "num_skills": 7,
    "categories": [
        "object", "object", "object", "object", "shape", "color",
        "size", "size"
    ],
    "skill": [
        "sheep", "highway", "sushi", "corgi", "heart", "yellow",
        "tiny", "tiny"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain sheep?  ",
        "Does the image contain a highway?  ",
        "Does the image contain sushi?  ",
        "Does the image contain a corgi?  ",
        "Is the highway heart-shaped?  ",
        "Is the color of the sheep yellow?  ",
        "Is the sheep tiny in size?  ",
        "Is the corgi tiny in size?"
    ],
    "scores": [
        1,
        0,
        1,
        1,
        0,
        1,
        1,
        1
    ]
}
```

657

Back to Table of Contents   30   Back to the First Page

Shapes Failure Case (Example 2, PixArt alpha)

Prompt: A huge, white, heart-shaped table is placed next to a chair.

Prompt Generation:

```
{
    "num_skills": 3,
    "categories": [
        "object", "size", "color", "shape"
    ],
    "skill": [
        "table", "huge", "white", "heart"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain a table?  ",
        "Is the table huge in size?  ",
        "Is the color of the table white?  ",
        "Is the shape of the table heart-shaped?"
    ],
    "scores": [
        1,
        0,
        0,
        0
    ]
}
```

658

Back to Table of Contents   31   Back to the First Page

659 **E.3 Sizes**

Sizes Failure Case (Example 1, DALL·E 3)

Prompt: In an oil painting, a tiny corgi is positioned in front of three tiny brown volcanoes.

Prompt Generation:

```
{
    "num_skills": 7,
    "categories": [
        "object", "object", "color", "style",
        "size", "number", "size", "spatial"
    ],
    "skill": [
        "corgi", "volcano", "brown", "oil painting", "tiny", "3",
        "tiny", "in front of"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain corgi?",
        "Does the image contain volcano?",
        "Is the color of the volcano brown?",
        "Is the style of the image oil painting?",
        "Is the size of the volcano tiny?",
        "Is the number of volcanoes exactly 3?",
        "Is the size of the corgi tiny?",
        "Is the corgi positioned in front of the volcano?"
    ],
    "scores": [
        1,
        1,
        1,
        1,
        0,
        0,
        0,
        1
    ]
}
```

Back to Table of Contents   32   Back to the First Page

660

Sizes Failure Case (Example 2, PixArt alpha)

Prompt: In an oil painting, a huge smartphone rests on a table next to a green corgi. A tiny hammer with a fluffy texture is also on the table, alongside a book.

Prompt Generation:

```
{
"num_skills": 7,
"categories": [
    "object", "object", "object", "size", "texture",
    "color", "size", "style"
],
"skill": [
    "smartphone", "corgi", "hammer", "huge", "fluffy",
    "green", "tiny", "oil painting"
]
}
```

Grading Results:

```
{
"questions": [
    "Does the image contain a smartphone?",
    "Does the image contain a corgi?",
    "Does the image contain a hammer?",
    "Is the smartphone huge in size?",
    "Is the hammer fluffy in texture?",
    "Is the corgi green in color?",
    "Is the hammer tiny in size?",
    "Is the style of the image oil painting?"
],
"scores": [
    1,
    1,
    1,
    0,
    0,
    0,
    1,
    1
]
}
```

661

Back to Table of Contents   33   Back to the First Page

662

E.4 Textures

Textures Failure Case (Example 1, PixArt alpha)

Prompt: A scene shows a glass-textured laptop on a desk beside a glass-textured robot. In the background, there is a duck standing on the floor next to a cactus.

Prompt Generation:

```
{
    "num_skills": 5,
    "categories": [
        "object", "object", "object", "object", "texture", "texture"
    ],
    "skill": [
        "laptop", "robot", "duck", "cactus", "glass", "glass"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain a laptop?  ",
        "Does the image contain a robot?  ",
        "Does the image contain a duck?  ",
        "Does the image contain a cactus?  ",
        "Does the robot have a glass texture?  ",
        "Does the laptop have a glass texture?"
    ],
    "scores": [
        1,
        1,
        1,
        1,
        0,
        1
    ]
}
```

663

Back to Table of Contents   34   Back to the First Page

Textures Failure Case (Example 2, DALL·E 3)

Prompt: In a vibrant countryside scene, a single wooden house stands in a field. Nearby, a corgi with a short tail observes a sheep grazing on the lush, green grass. In the background, a fluffy-textured volcano looms under a clear blue sky. On a wooden bench beside the house, a yellow screwdriver lies next to a metal hammer.

Prompt Generation:

```
{
    "num_skills": 7,
    "categories": [
        "object", "object", "object", "object", "object",
        "object", "color", "texture"
    ],
    "skill": [
        "house", "corgi", "sheep", "volcano", "screwdriver", "hammer",
        "yellow", "fluffy"
    ]
}
```

Grading Results:

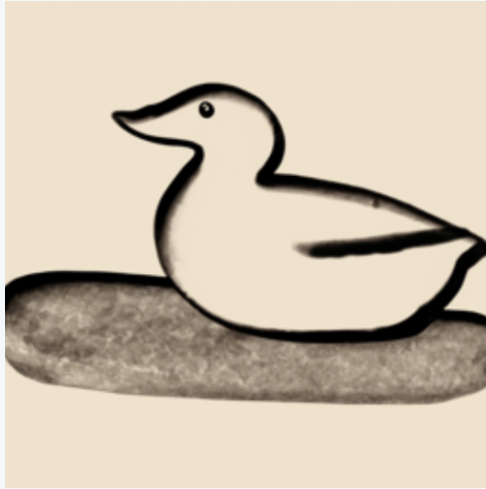
```
{
    "questions": [
        "Does the image contain a house?   ",
        "Does the image contain a corgi?   ",
        "Does the image contain a sheep?   ",
        "Does the image contain a volcano?   ",
        "Does the image contain a screwdriver?   ",
        "Does the image contain a hammer?   ",
        "Is the color of the screwdriver yellow?   ",
        "Does the volcano have a fluffy texture?"
    ],
    "scores": [
        1,
        1,
        1,
        1,
        1,
        1,
        1,
        0
    ]
}
```

664

Back to Table of Contents   35   Back to the First Page

665 **E.5 Spatial Relationship**

Spatial Failure Case (Example 1, DeepFloyd IF XL v1)

Prompt: A tiny glass-textured duck is positioned on the right side of a rock in an ink-style image.

Prompt Generation:

```
{
    "num_skills": 4,
    "categories": [
        "object", "size", "texture", "style", "spatial"
    ],
    "skill": [
        "duck", "tiny", "glass", "ink", "right"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain a duck?",
        "Is the size of the duck tiny?",
        "Does the duck have a glass texture?",
        "Is the style of the image ink?",
        "Is the duck positioned on the right side of the rock?"
    ],
    "scores": [
        1,
        0,
        0,
        1,
        0
    ]
}
```

666

Back to Table of Contents   36   Back to the First Page

Spatial Failure Case (Example 2, PixArt alpha)

Prompt: The image shows four white, triangle-shaped pine trees with a fluffy texture. A rock is positioned at the bottom of each pine tree, touching them.

Prompt Generation:

```
{
    "num_skills": 5,
    "categories": [
        "object", "shape", "color", "texture", "number", "spatial"
    ],
    "skill": [
        "pine tree", "triangle", "white", "fluffy", "4", "bottom"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain pine trees?  ",
        "Are the pine trees triangle shaped?  ",
        "Are the pine trees white in color?  ",
        "Do the pine trees have a fluffy texture?  ",
        "Is the number of pine trees exactly four?  ",
        "Is a rock positioned at the bottom of each pine tree,
        touching them?"
    ],
    "scores": [
        1,
        1,
        0,
        1,
        0,
        0
    ]
}
```

667

Back to Table of Contents   37   Back to the First Page

E.6   Styles

Styles Failure Case (Example 1, SD v1.4)

Prompt: A brown duck in an expressionist style.

Prompt Generation:

```
{
    "num_skills": 2,
    "categories": [
        "object", "color", "style"
    ],
    "skill": [
        "duck", "brown", "expressionism"
    ],
    "question": [
        "Does the image contain a duck?  ",
        "Is the duck brown?  ",
        "Is the style of the image expressionism?"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain a duck?  ",
        "Is the duck brown?  ",
        "Is the style of the image expressionism?"
    ],
    "scores": [
        1,
        1,
        0
    ]
}
```

Back to Table of Contents   38   Back to the First Page

Styles Failure Case (Example 2, SD v2.1)

Prompt: A huge fork is positioned nearer to the observer than a plate in an impressionism-style image.

Prompt Generation:

```
{
    "num_skills": 3,
    "categories": [
        "object", "style", "size", "spatial"
    ],
    "skill": [
        "fork", "impressionism", "huge", "in front of"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain a fork?",
        "Is the style of the image impressionism?",
        "Is the fork huge?",
        "Is the fork positioned nearer to the observer or camera than
        the plate?"
    ],
    "scores": [
        1,
        0,
        0,
        0
    ]
}
```

670

Back to Table of Contents   39   Back to the First Page

671

E.7 Colors

Colors Failure Case (Example 1, DALL·E 3)

Prompt: The image shows a green cow standing beside a tiny truck. There is a hammer placed on the ground near them, and a large bicycle is parked in the background.

Prompt Generation:

```
{
    "num_skills": 4,
    "categories": [
        "object", "object", "object", "size", "color"
    ],
    "skill": [
        "hammer", "truck", "cow", "tiny", "green"
    ]
}
```

Grading Results:

```
{
    "questions": [
        "Does the image contain a hammer?  ",
        "Does the image contain a truck?  ",
        "Does the image contain a cow?  ",
        "Is the truck tiny?  ",
        "Is the cow green?"
    ],
    "scores": [
        1,
        1,
        1,
        1,
        0
    ]
}
```

672

Back to Table of Contents   40   Back to the First Page

Colors Failure Case (Example 2, PixArt alpha)

Prompt: The graffiti-style image features a gray cat and a zebra.

Prompt Generation:

```
{
    "num_skills": 3,
    "categories": [
        "object", "object", "color", "style"
    ],
    "skill": [
        "zebra", "cat", "gray", "graffiti"
    ]
}
```

Grading Results:

```
{
"questions": [
    "Does the image contain a zebra?",
    "Does the image contain a cat?",
    "Is the color of the cat gray?",
    "Is the style of the image graffiti?"
],
"scores": [
    0,
    1,
    0,
    1
]
}
```

673

Back to Table of Contents   41   Back to the First Page