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

A.3 Results

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

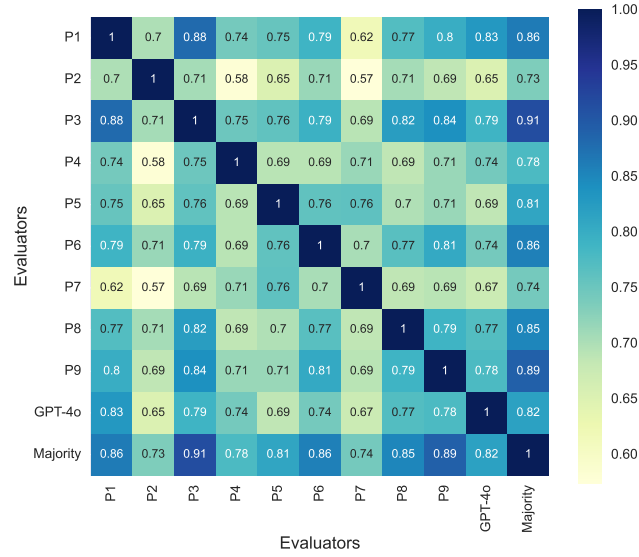



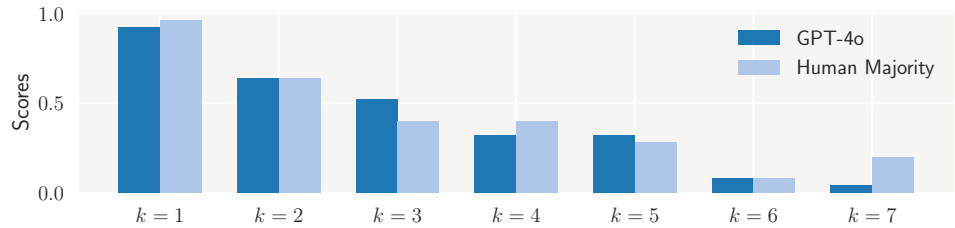
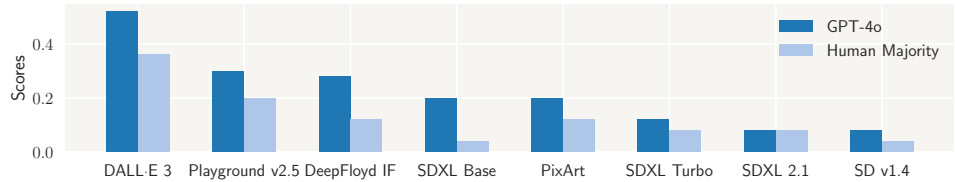
Figure 9: **Pairwise Consistency Heatmap.** The heatmap shows the consistency between different human evaluators (**P1** to **P9**) as well as a majority vote (**Majority**) and GPT-4o (**GPT-4o**) across all k for DALL·E 3. Each cell represents the consistency score, with darker shades showing higher 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.

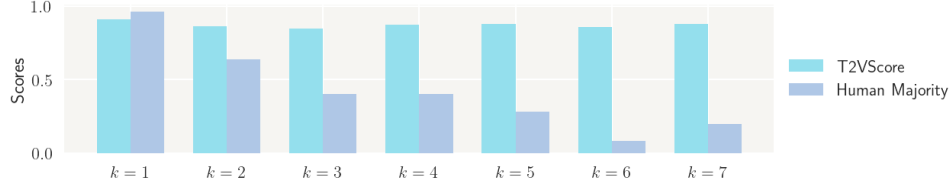
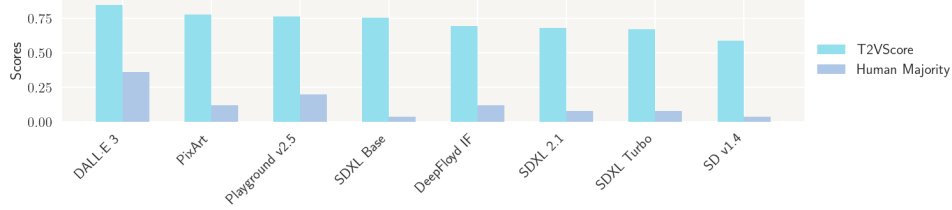
Back to Table of Contents   16   Back to the First Page

(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.

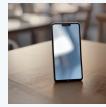
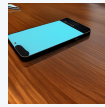
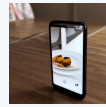
Compare with Prior Grading Approach. We further conduct experiments with previous state-of-the-art grading approach [28] and compare them with human preferences. As shown in Fig. 10 and Fig. 11, our grading method aligns better with human preferences, for example, in Fig. 10a, as k grows, both our grading 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
```

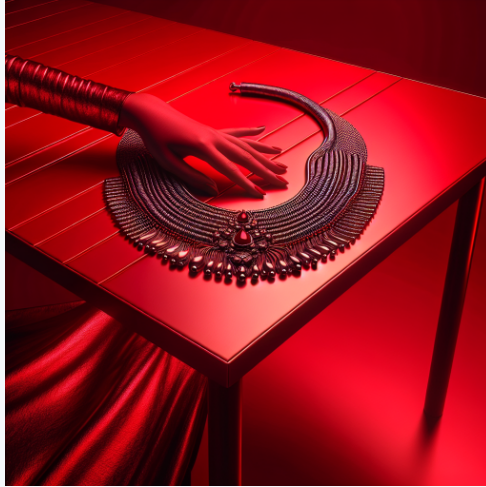
Automatic grading questions:

```
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
```

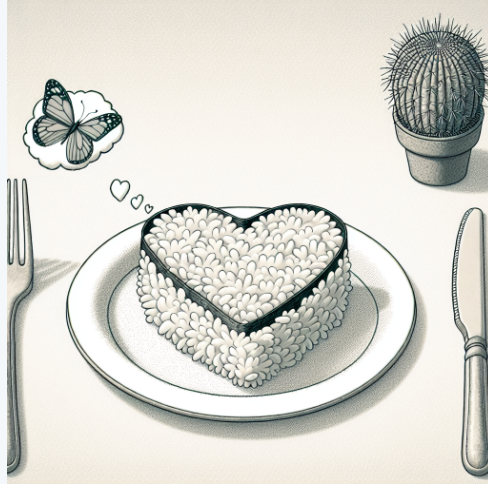
GPT-4o grading details:

```
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:

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.

617
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

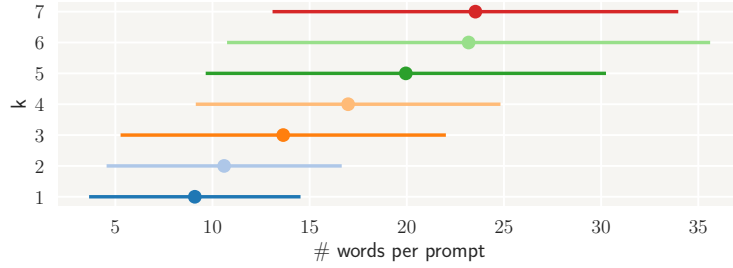



Figure 12: Distribution of prompt length in CONCEPTMIX: Larger values of k result in longer and potentially more complex prompts.

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

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

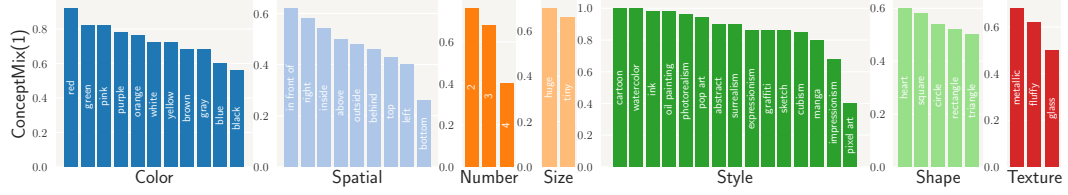



Figure 13: **Performance of concepts within the same category.**

Tab. 7 provides the concept fraction score of all evaluated models, showing a high correlation with the full mark score reported in Tab. 3. Similar to Tab. 3, the concept fraction score drops when increasing k, with DALL·E 3 being the best, and SD v1.4 being the worst. Note the drop in concept fraction score not only indicates the difficulty level increase of the whole text prompts but also shows each 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]	$0.74_{\pm0.03}$	$0.61_{\pm0.03}$	$0.55_{\pm0.03}$	$0.50_{\pm0.02}$	$0.44_{\pm0.02}$	$0.41_{\pm0.02}$	$0.36_{\pm0.02}$
SD v2.1 [33]	$0.74_{\pm0.03}$	$0.68_{\pm0.03}$	$0.61_{\pm0.03}$	$0.54_{\pm0.03}$	$0.50_{\pm0.03}$	$0.48_{\pm0.02}$	$0.45_{\pm0.02}$
SDXL Turbo [39]	$0.81_{\pm0.03}$	$0.72_{\pm0.03}$	$0.65_{\pm0.03}$	$0.60_{\pm0.03}$	$0.57_{\pm0.02}$	$0.54_{\pm0.02}$	$0.49_{\pm0.02}$
PixArt alpha [7]	$0.82_{\pm0.03}$	$0.73_{\pm0.03}$	$0.67_{\pm0.03}$	$0.61_{\pm0.03}$	$0.56_{\pm0.02}$	$0.53_{\pm0.02}$	$0.49_{\pm0.02}$
SDXL Base [33]	$0.84_{\pm0.03}$	$0.76_{\pm0.03}$	$0.69_{\pm0.02}$	$0.63_{\pm0.02}$	$0.60_{\pm0.02}$	$0.57_{\pm0.02}$	$0.53_{\pm0.02}$
DeepFloyd IF XL v1 [43]	$0.84_{\pm0.03}$	$0.74_{\pm0.03}$	$0.66_{\pm0.03}$	$0.61_{\pm0.02}$	$0.59_{\pm0.02}$	$0.55_{\pm0.02}$	$0.51_{\pm0.02}$
Playground v2.5 [26]	$0.84_{\pm0.03}$	$0.77_{\pm0.03}$	$0.71_{\pm0.02}$	$0.64_{\pm0.02}$	$0.62_{\pm0.02}$	$0.58_{\pm0.02}$	$0.52_{\pm0.02}$
DALL·E 3 [2]	$0.92_{\pm0.02}$	$0.85_{\pm0.02}$	$0.83_{\pm0.02}$	$0.76_{\pm0.02}$	$0.75_{\pm0.02}$	$0.72_{\pm0.02}$	$0.71_{\pm0.02}$

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

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
    ]
}
```

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
    ]
}
```

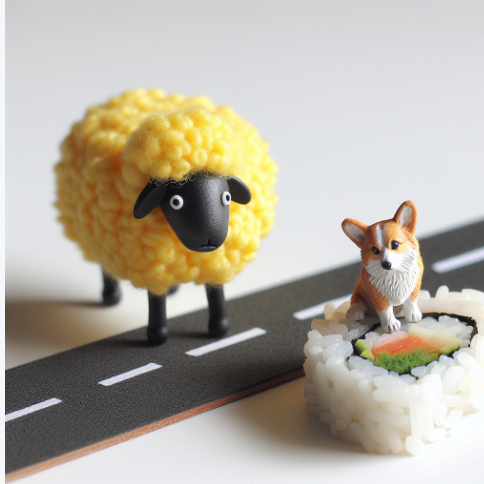
655

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**

<div style="border:1px solid #2b7bb9; border-radius:8px;">

<div style="background:#2b7bb9; color:white; text-align:center;">Sizes Failure Case (Example 1, DALL·E 3)</div>

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
    ]
}
```

</div>

660

Back to Table of Contents   32   Back to the First Page

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
    ]
    }
```

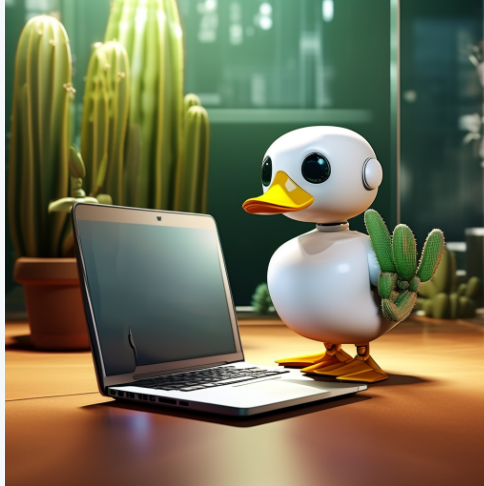
661

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
    ]
}
```

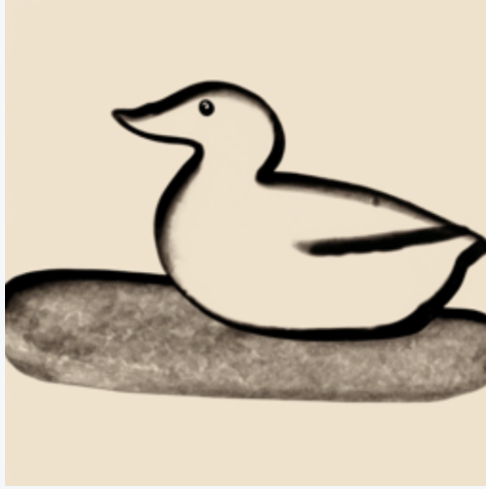
664

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

668 **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
    ]
}
```

669

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**

<div style="border:1px solid #2878b5; border-radius:8px;">

<div style="background:#2878b5; color:white; text-align:center;">Colors Failure Case (Example 1, DALL·E 3)</div>

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
    ]
}
```

</div>

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