Data Card & Data Sheet for CONCEPTMIX

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

Princeton University

Abstract

In the supplementary materials, we document the benchmark & dataset details of CONCEPTMIX for research transparency. More technical details can be found in the appendix of the main paper. Our code & website are available here: `https://github.com/princetonvisualai/ConceptMix`.

1 Data Card   **2**

   1.1   Data Details 2

   1.2   Prompt Generation 2

   1.3   Question Generation 4

   1.4   Author Statement 4

   1.5   URL for Benchmark Access and Metadata 4

2 Data Sheet   **5**

   2.1   Motivation 5

   2.2   Composition / Collection process / Preprocessing / Cleaning 5

   2.3   Distribution 5

   2.4   Maintenance 5

---

[*]Equal contribution

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

1 Data Card

1.1 Data 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 $k = 1$. Note that we use all objects for both easy and hard splits to ensure a fair comparison.

1.2 Prompt Generation

We use GPT-4o (endpoint of May 13th, 2024), to help bind multiple concepts and generate prompts. 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".

2

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.

Here the property description of each selected concept category is generated using the template provided in Table 1.

Table 1: 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, bottom	[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]

After generating the prompts, we then prompt GPT-4o for validation, using the following 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.

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

Prompt length. We also provide the distribution of text prompt lengths for different values of k. The length of the text prompt may indicate the complexity of the task, as longer prompts tend to involve more concepts. The distribution of text prompt lengths for each k is shown in Figure 1.

3
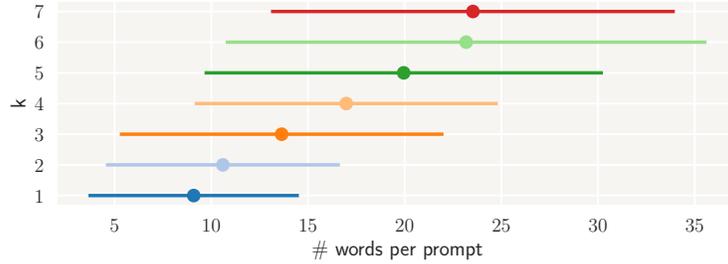
Figure 1: Distribution of prompt length in CONCEPTMIX: Larger values of k result in longer and potentially more complex prompts.

1.3 Question Generation

For each generated prompt, we also accompany it with a list of GPT-4o-generated questions, 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.

1.4 Author Statement

We, the authors, bear all responsibility in case of violation of rights. We confirm that all data and content included in this paper have been obtained and processed in compliance with relevant ethical guidelines.

1.5 URL for Benchmark Access and Metadata

Benchmark Access: https://github.com/princetonvisualai/ConceptMix & https://huggingface.co/spaces/xindiw/ConceptMix
Croissant Metadata Record: https://huggingface.co/spaces/xindiw/ConceptMix/blob/main/croissant_metadata.json

4

2 Data Sheet

In addition to the data card, we provide a data sheet following the documentation frameworks provided by [1]. Note that we do not release a fixed dataset, but the code for generating the dataset and grading.

2.1 Motivation

For what purpose was the dataset created?

Our benchmark is designed to evaluate the capability of text-to-image models to generate images containing a random combination of visual concepts. Our benchmark focuses on the alignment between the generated image and given prompt, rather than other aspects of the image quality (e.g., aesthetics).

2.2 Composition / Collection process / Preprocessing / Cleaning

The answers are described in the data card (Section 1).

2.3 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

No. Our benchmark will be managed and maintained by the author group.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)?

The code will be released to the public and hosted on GitHub and Huggingface.

When will the dataset be distributed?

We will release the code of our benchmark in June 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The code of our benchmark will be distributed under the MIT license.

2.4 Maintenance

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Please contact Xindi Wu (`xindiw@princeton.edu`), who is responsible for maintenance.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes, if we find any errors in the generation and grading pipeline, we'll update the code and revise the results accordingly. We'll also announce any changes on our website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

For benchmark contributions, the most efficient way to reach us is via GitHub pull requests. For more questions, please contact Xindi Wu (`xindiw@princeton.edu`), who is responsible for maintenance.

5

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

6