ConceptMix: A Compositional Image Generation Benchmark with Controllable Difficulty

Xindi Wu*[1], Dingli Yu*[12], Yangsibo Huang*[13], Olga Russakovsky[1], Sanjeev Arora[1]

[1]Princeton University, [2]Microsoft Research, [3]Google Research

* Equal contribution

NEURAL INFORMATION PROCESSING SYSTEMS

Website | Paper | Code

Compositionality in T2I Generation

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

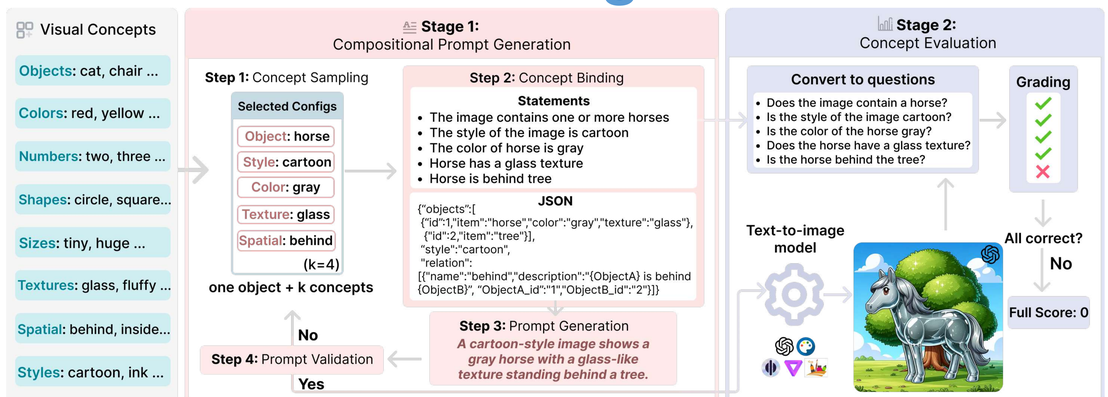
- Rely heavily on human-designed text prompts or fixed templates, limiting their diversity and complexity
- Evaluations have **low discriminative power**

Benchmark	# Concept Categories	# Concepts Per Text Prompt	Concept Binding Approach
CC-500 [1]	2	2	Fixed template
ABC-6K [1]	2	2	Fixed template
Attn-Exct [2]	4	2	Fixed template
HRS-comp [3]	2	≤ 3	Fixed template
T2I-CompBench [4]	6	≤ 5	Fixed template, GPT augmented
ConceptMix (ours)	8	Unlimited	Free-form, GPT-4o generated

How to automatically generate prompts that compose diverse visual concepts?

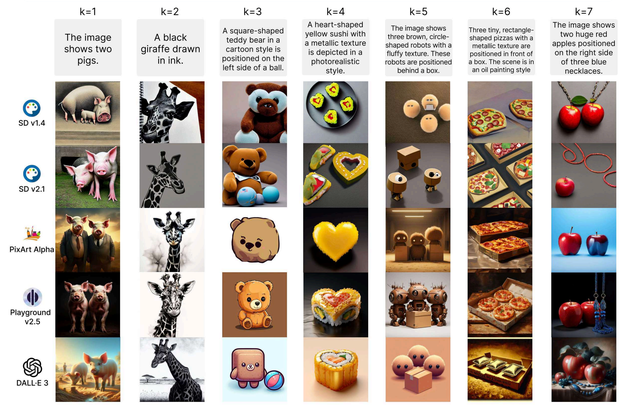
How to automatically grade results based on (prompt, generated image)?

How to evaluate compositionality of Text-to-Image models?



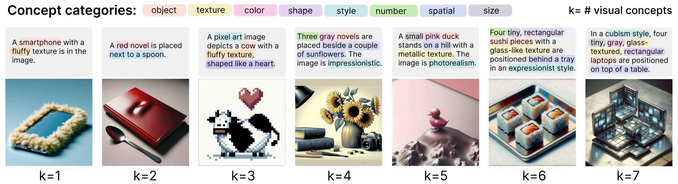
We introduce ConceptMix, a scalable benchmark that evaluates T2I models' compositionality with controllable difficulty, using GPT-4o to generate prompts and grade the images.

Qualitative Results



As prompts become more complex, image quality degrade. DALL·E 3 performs best, while SD v1.4 performs worst.

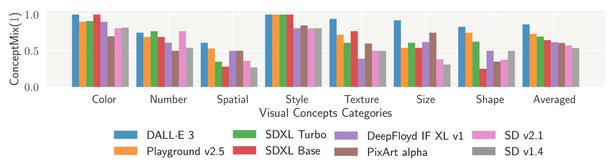
ConceptMix



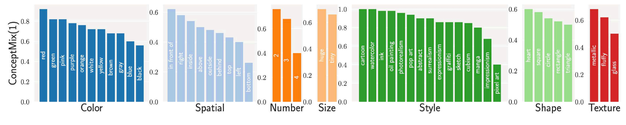
Two-stage approach:

- **Compositional Prompt Generation**: ConceptMix uses GPT-4o to create diverse and complex prompts by combining one object with k random visual concepts.
- **Concept Evaluation**: ConceptMix computes concept accuracy by having GPT-4 answer concept-specific questions.

Individual Concept Performance (k=1)



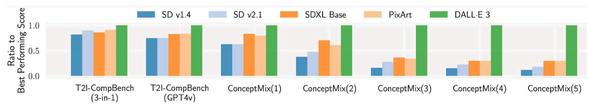
Takeaway: Color and style are the easiest while spatial, size, and shape are challenging.



Takeaway: Varying concept performance are observed within the same concept category.

Compositional Generation (k > 1)

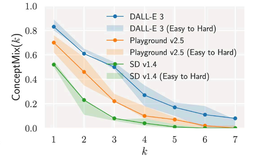
- ConceptMix Shows Stronger Discriminative Power



- Performance of Eight T2I Models on ConceptMix

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
SD v1.4	0.52	0.23	0.08	0.03	0.01	0.00	0.00
SD v2.1	0.52	0.29	0.14	0.06	0.03	0.01	0.00
SDXL Turbo	0.64	0.35	0.18	0.09	0.03	0.02	0.01
PixArt alpha	0.66	0.37	0.17	0.09	0.05	0.01	0.01
DeepFloyd IF XL v1	0.68	0.38	0.21	0.09	0.05	0.02	0.01
SDXL Base	0.69	0.43	0.18	0.09	0.05	0.01	0.00
Playground v2.5	0.70	0.46	0.22	0.10	0.07	0.02	0.00
DALL·E 3	0.83	0.61	0.50	0.27	0.17	0.11	0.08

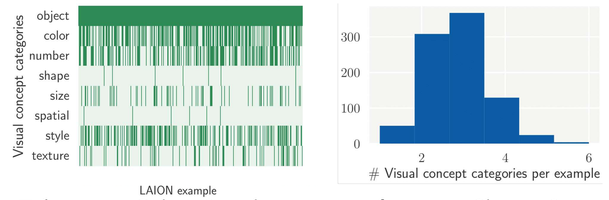
Takeaway: We observe a consistent performance drop as k increases with the leading proprietary model, DALL·E 3, struggling at k = 5.



Training Data Concept Diversity

Why are models bad at k ≥ 3?

- Limited exposure to complex concept combinations
- Disparate concept representation



Takeaway: Colors & styles are most frequent; Shapes & spatial relationships are least; Most examples include 2-3 concepts.

More details (in paper):
Human evaluation & more experiment results

References:

[1] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. NeurIPS 2023

[2] 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. TOG 2023

[3] 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. ICCV 2023

[4] Huang, Kaiyi, et al. "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation." NeurIPS 2023.