

ScImage: How Good are Multimodal Large Language Models at Scientific Text-to-Image Generation?

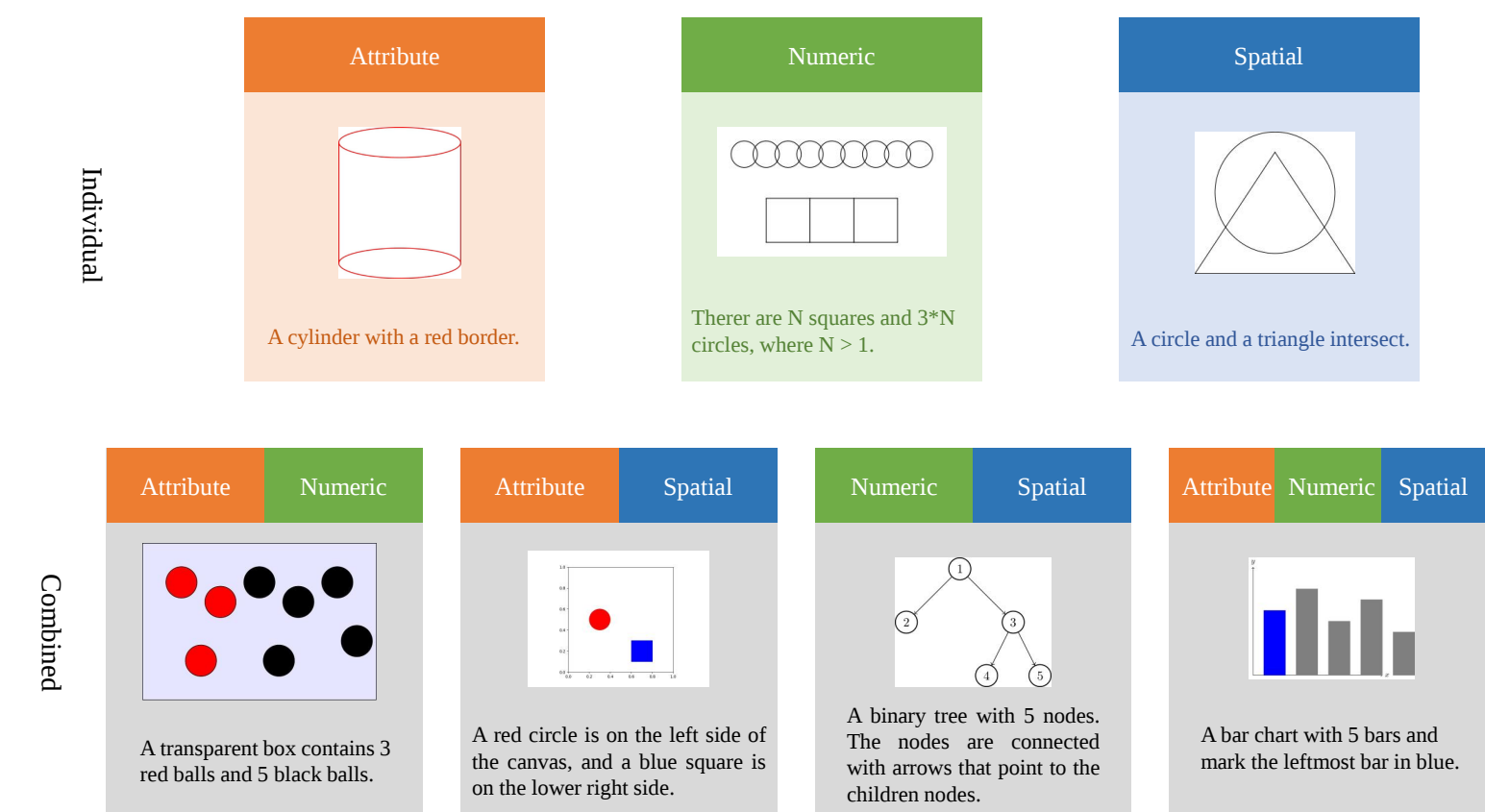
Authors: Leixin Zhang, Steffen Eger, Yinjie Cheng, Weihe Zhai, Jonas Belouadi, Fahimeh Moafian, Zhixue Zhao

Overview

- **Background:** Multimodal large language models (LLMs) have demonstrated impressive capabilities in text-to-image generation. However, LLMs’ performance in generating **scientific images** remains underexplored.
- **Benchmark:** 404 items for testing model capability of scientific text-to-image generation.
- **Dimensions:** Image generation assessment on three understanding dimensions—numeric, spatial, and attribute comprehension.
- **Models:** Testing on ScImagewith seven state-of-the-art models: GPT-4o, GPTo1, Llama-3.1, Automatikz, Qwen-2.5, DALL-E and Stable Diffusion.
- **Two output modes:** code-based (text-code-image) and genuine multimodal (text-image) settings.

Understanding Dimensions

- **Attribute:** color, shape, size, length, etc.
- **Numeric:** number, percentage, etc.
- **Spatial:** "intersect", "to the left", "bottom", etc.



Output Modes

- **Direct Text-Image**

Models: AutomaTikZ and Dall-E

Prompt:Please generate a scientific figure according to the following requirements:
{generation query}

- **Text-Code-Image** (Python & Tikz)

Models: GPT-4o, Llama-3.1-8B, AutomaTikZ, GPT-o1 and Qwen2.5-Coder-7B-Instruct

Prompt: Please generate a scientific figure according to the following requirements:
{generation query}. Your output should be in [Python/Tikz] code. Do not include any text other than the [Python/Tikz] code.

Human Evaluation

- **11 Expert Annotators:** PhD students and higher, differentiated from crowd workers.
- **Three Criteria:** Evaluate generated images on correctness, relevance, and scientificness.

Model Performance on English Prompts

- GPT-4o outperforms all other models and achieves the highest scores for both the text-tikz-image and text-python-image output.

Model	Correctness	Relevance	Scientific style	Compile Error Rate
Automatikz	2.05	2.31	3.35	0.04
Llama_tikz	1.78	1.94	2.61	0.29
GPT-4o_tikz	3.50	3.67	3.75	0.09
Llama_python	2.10	2.54	3.18	0.28
GPT-4o_python	3.51	3.40	3.93	0.07
Stable Diffusion	2.19	2.09	1.96	-
DALL-E	2.16	2.00	1.55	-

Table: Overall model performance, averaged across instances and annotators.

Model Performance on Multilingual Prompts

Criteria	Correctness				Relevance				Scientific style			
	EN	DE	ZH	FA	EN	DE	ZH	FA	EN	DE	ZH	FA
Language												
Llama_tikz	1.88	1.48	1.50	1.23	2.18	1.78	2.10	1.68	2.78	2.23	2.80	2.90
GPT-4o_tikz	3.85	4.03	3.98	3.68	4.03	4.23	4.60	3.98	4.10	4.43	4.40	3.98
OpenAI-o1_tikz	4.43	3.68	3.83	4.05	4.45	3.80	4.10	4.18	4.40	3.88	4.03	4.05
Llama_python	2.53	1.35	1.75	1.78	2.70	1.53	2.00	1.90	3.20	2.50	3.10	3.30
GPT-4o_python	3.38	4.15	4.13	3.48	3.35	4.18	4.23	3.35	3.88	4.50	4.83	3.85
OpenAI-o1_python	4.28	3.45	4.10	3.60	4.10	3.45	3.93	3.60	4.50	4.08	4.30	4.05
Qwen2.5_python	3.10	2.30	2.05	2.40	3.08	2.48	2.25	2.53	3.70	3.43	3.28	3.68
DALL-E	1.98	2.15	1.83	1.93	1.88	2.03	2.03	2.00	1.40	1.58	1.53	1.50
Average	3.18	2.82	2.89	2.77	3.22	2.93	3.15	2.90	3.49	3.33	3.53	3.41

Table: Multilingual performance of overall generations. The best is highlighted in bold per metrics per language.

Results of Understanding Dimensions

- Numerical comprehension poses the greatest challenge.
- Due to models’ weakness in spatial understanding, tasks that involve combined understanding types also tend to receive lower scores.

Model	Attribute	Numerical	Spatial	Attribute & Numerical	Attribute & Spatial	Numerical & Spatial	Attribute & Numerical & Spatial
Automatikz	2.42	1.91	1.71	2.29	2.04	2.13	1.77
Llama_tikz	2.53	1.55	1.77	1.69	1.84	1.91	1.3
GPT-4o_tikz	4.11	3.49	3.35	3.41	3.53	3.59	3.13
Llama_python	2.24	2.38	1.96	2.28	2.28	1.63	1.97
GPT-4o_python	3.95	3.92	3.47	3.46	3.34	3.28	3.13
Stable Diffusion	2.75	1.73	2.06	2.41	2.46	1.96	2.11
DALL ·E	2.68	1.77	2.13	2.36	2.31	1.94	2.07
Average	2.95	2.39	2.35	2.56	2.54	2.35	2.21
Sample Size	48	64	56	80	40	64	52

Table: Correctness evaluation within each understanding category.



UNIVERSITY
OF TWENTE.

