Description Logics for T2I Evaluations

Vihang Pancholi **Harsh Panchal** **Aishwariya Ranjan** **Aditi Ganapathi**

1 Introduction

The dynamic combination of computer vision and natural language processing has led to a thriving field of text-to-image synthesis research. Although existing models show promise, the standard of the input prompts strongly influences the accuracy and pertinence of the generated visual outputs.

The primary obstacle is the complex comprehension of textual descriptions, which calls for a change from traditional prompt generating techniques to a more advanced model. With the help of description logics, which give a formalism for capturing and deducing the semantics of natural language, specificity and context coherence can be strengthened rapidly. In the context of text-to-image generative models, automated assessment frameworks are a critical requirement. This study aims to meet this need by closely examining the performance of the Stable Diffusion V1.4 and Stable Diffusion V2.1 models. Though these models work well overall, there are some situations when they fail, which poses serious problems that need to be carefully investigated and resolved.

When assessment data contains biases, like in the case of assumptions like "Apples are always red or green," questions are raised regarding how the model learns. The goal of this study is to evaluate the viability of developing a more impartial evaluation dataset to reduce biases and determine how these biases affect the performance of the Stable Diffusion model. The evaluation method must take into consideration factors other than color, such as the fact that apples are rarely compared to sizes. The study investigates how flexible Stable Diffusion is in producing apple images at non-traditional scales, raising important concerns regarding the model's intrinsic biases and ability to produce a variety of outputs. Evaluation becomes more challenging when T2I models start to exhibit hallucinations, in which expected pictures are replaced by unexpected outputs, such as apple pies instead of apples.

This project addresses another particular and crucial part of text-to-image synthesis—evaluating the quality of generated visuals—in addition to the basic problem of prompt creation. Using this approach, we want to analyze and find gaps in text-to-image models' performance and allow them to produce images that are more closely aligned with the complex nuances of human language.

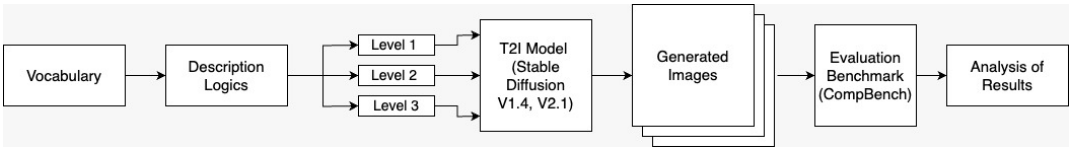

Figure 1: Project pipeline

2 Approach

2.1 Description Logics (DL)

In the larger subject of artificial intelligence and knowledge representation, DL is a formalism that seeks to represent and reason on the relationships and properties of objects in a domain. With its foundation in mathematical logic, DL offers a systematic and expressive way to describe knowledge,

which makes it a vital component of many different applications, including knowledge-based systems, ontology engineering, and semantic web technologies.

2.2 Prompt Generation

A vocabulary that included objects, properties, and grammatical elements like prepositions was created for our study. The appropriate object-property mappings were identified. Then, using DL rules, we systematically created prompts in our T2I model evaluation technique to evaluate the performance of the model at various levels of complexity. For Level 1, we generated fundamental prompts using DL rules by arranging an object selected randomly, a property, and an article. The format was expanded to contain an article, property, object, preposition, and another object at Level 2, which resulted in an increase in complexity. These were generated by adding prepositions at random to Level 1 prompts. Level 3 prompts were created by combining two randomised Level 1 prompts with a preposition.

For instance, consider these sample sets and their corresponding prompts per level.

$C = Color = \{Red,\ Green,\ Black\}$

$D = Fruit = \{Banana,\ Apple\}$

$F = Furniture = \{Chair,\ Table\}$

$R = Relation = \{\text{“}on\ top\ of\text{”},\ \text{“}and\text{”}\}$

Level 1 $A \cup B$ = {“$black\ apple$”, “$red\ banana$”}

Level 2 $R((C \cup D), F)$ = {“$black\ apple\ on\ top\ of\ chair$”}

Level 3 $R((C \cup D), (C \cup D))$ = {“$black\ apple\ and\ red\ banana$”}

2.3 Image Generation

The next step was to generate images using the prompts generated in the previous step. Taking use of the extensive availability and open-source nature of Stable Diffusion v1.4 and v2.1, these models were used to produce 500 photos for each level. Stable Diffusion models work by iteratively fine-tuning a random noise vector until the noise matches the desired text prompts. This produces high-quality images. By means of a sequence of denoising stages, the diffusion process facilitates the regulated conversion of the noise vector into visually consistent and contextually appropriate outputs that correspond to the designated textual descriptions. Because these processes need a lot of processing power, Google Colab's T4 GPUs, Kaggle Notebooks' T4 GPUs, and the ASU supercomputer Sol were used to increase processing speed. During this phase, 3000 pictures were generated in total. Each image file was saved with the nomenclature suggested in the benchmark documentation.

2.4 Application of Evaluation Benchmark

T2I-CompBench stands as an invaluable resource for users assessing text-to-image models. Following the creation of images for every level and model, an assessment was carried out using the T2I-CompBench model to evaluate attribute binding in the images that were generated. Each image was distinguished from the others by the prompt that was included in its name. Questions were generated by methodically extracting prompts from the image names, each of which contained a single object-property pair.

3 Results

3.1 Prompt Generation

Using the lexicon that was carefully created for this study, prompts that adhered to the hierarchical structures defined at each of the three levels above were produced in an organized manner. A careful

Table 1: Sample prompts generated for each level

Level 1	Level 2	Level 3
a short envelope	two peach to the right of orange	a square Burj Khalifa on top of one doctor
a pink desert	an orange forest far away from grater	a yellow okra near a narrow Venus
three ice cream	two Meryl Streep in front of scissors	a green school on top of a green ambulance
a yellow whisk	a green bicycle on fall	a purple post office behind an orange lettuce
an orange rose	an orange Sydney Opera House in between kiwi	a narrow cricket match beside a big mushroom

selection process resulted in the random curation of 500 prompts at each level in the context of performing a comparative analysis between the two diffusion models. A sample of prompts from each level can be found in the Table 1.

3.2 Image Generation

Figures 2, 3, and 4 show the images generated by Stable Diffusion v1.4 and v2.1 for each level. One is able to draw a comparison between the performance of the models upon first glance.

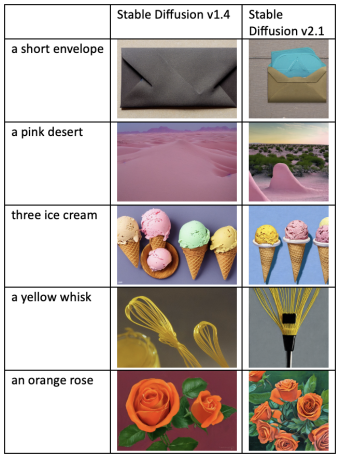

Figure 2: Level 1 images

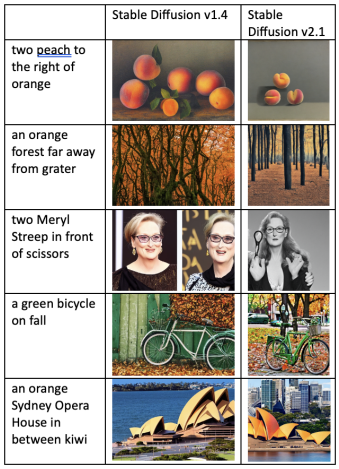

Figure 3: Level 2 images

Figure 4: Level 3 images

3.3 Application of Evaluation Benchmark

Our text-to-image generative models underwent a thorough evaluation as part of the T2I-CompBench benchmark implementation. The resulting findings, which are contained in a JSON file, capture the complex relationship between the models' responses and the given prompts. Every question in this file has its own question ID, which creates a traceable connection between the prompts and their associated results. The corresponding probabilities, which indicate the chance of getting an answer in the affirmative for each question, provide specific details about how accurate the model is at predicting outcomes.

4 Analysis and Conclusion

After closely examining the scatterplots for each unique prompt complexity level for Stable Diffusion v1.4 as shown in Figures 5, 6, and 7, and for Stable Diffusion v2.1 in Figures 8, 9, and 10, a recognizable and fascinating collection of patterns emerges. Interestingly, a chance of "yes" equaling 1 is prominently frequent in Level 1, indicating a strong inclination towards positive answers. But as prompt complexity increases, this pattern gradually becomes less pronounced, as shown in the scatterplots for Levels 2 and 3. Further insights are provided by the calculated mean and median

Table 2: Comparison of statistical measures

Statistical Measure	Stable Diffusion v1.4			Stable Diffusion v2.1		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Mean	0.759	0.286	0.238	0.816	0.313	0.308
Median	0.880	0.157	0.123	0.912	0.237	0.247
Min	0.006	0.000	0.000	0.009	0.000	0.000
Max	1.000	0.995	0.955	1.000	0.994	0.959

probability, which show a declining trend with increasing prompt complexity. The statistical analysis reveals a significant decline in the mean and median probabilities between Level 1 and Level 2, which is an intuitive but unfavourable observation. These comparisons can be viewed in Table 2.

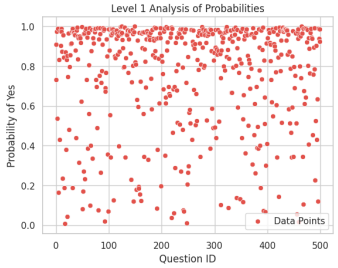

Figure 5: Level 1 scatterplot SD1.4

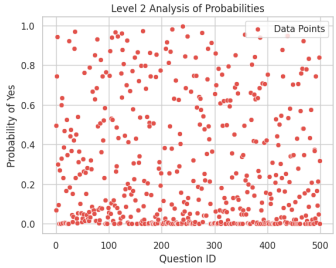

Figure 6: Level 2 scatterplot SD1.4

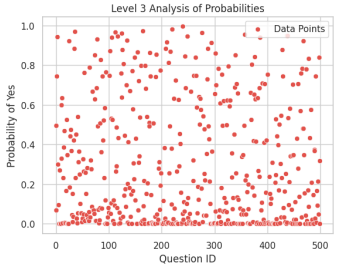

Figure 7: Level 3 scatterplot SD1.4

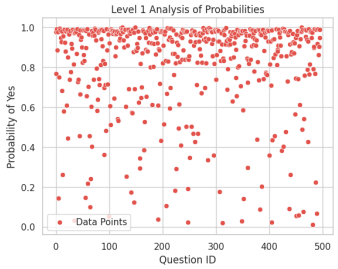

Figure 8: Level 1 scatterplot SD2.1

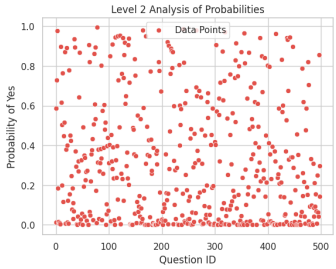

Figure 9: Level 2 scatterplot SD2.1

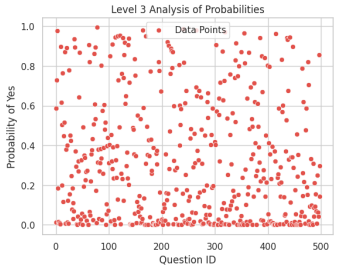

Figure 10: Level 3 scatterplot SD2.1

5 Individual Contribution

Vihang Pancholi Literature review(DL, Evaluation Benchmarks), vocabulary formation, running Stable Diffusion v1.4 for image generation, CompBench evaluation for V1.4 images, result analysis, presentation

Aditi Ganapathi Literature review(DL, Evaluation benchmarks), vocabulary formation, prompt generation for Levels 1, 2 ,3, CompBench evaluation for V2.1 images, result analysis, report

Aishwariya Ranjan Literature review (DL, T2I models), vocabulary formation, prompt generation for Levels 1, 2 ,3, report

Harsh Panchal Literature review (DL, T2I models), vocabulary formation, running Stable Diffusion v2.1 for image generation, presentation

References

[1] Huang, K., Sun, K., Xie, E., Li, Z., Liu, X. (2023). T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. arXiv preprint arXiv:2307.06350.

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

[3] Cho, J., Zala, A., Bansal, M. (2023). Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3043-3054).

[4] Bakr, E. M., Sun, P., Shen, X., Khan, F. F., Li, L. E., Elhoseiny, M. (2023). Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20041-20053).

[5] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H. (2016, June). Generative adversarial text to image synthesis. In International conference on machine learning (pp. 1060-1069). PMLR.