Investigation of the quality of svg generation by the Qwen3 family of models

Kuzin Aleksey, Artyom Zaitsev

June 2025

Abstract

In this paper, we investigate the ability of the Qwen3-0.6B,Qwen3-8b, and Qwen3-14b models to generate svg based on a text description. We use two benchmarks, one to evaluate the aesthetic scores of images and the second to match the image with a given text. Project link: `https://github.com/teodor-r/TextToSvg`.

1 Introduction

The ability of large language models (LLMs) to generate SVG (Scalable Vector Graphics) code based on a given textual description represents a unique intersection between natural language understanding , symbolic reasoning , and visual representation generation . This task is challenging overall. The ability of LLMs to generate SVG code has practical and theoretical importance.

- **Bridging Language and Visual Representation:** This capability demonstrates that language models can extend beyond text generation to influence visual design, enabling new applications in UI/UX prototyping, education, and creative tools.

- **Low-Code / No-Code Applications:** Non-technical users could describe an image or diagram and receive valid SVG output, lowering the barrier to entry for creating scalable, editable vector graphics.

- **Integration with Web Technologies:** Since SVG is natively supported by web browsers and integrates seamlessly with HTML/CSS/JS, this skill empowers models to contribute directly to web development workflows.

- **Evaluation of Reasoning Capabilities:** SVG generation serves as a benchmark for structured, multimodal reasoning — testing not only language understanding but also spatial logic, compositional generalization, and syntax-aware generation.

- **Foundation for More Complex Tasks:** Mastering SVG generation can pave the way for more advanced tasks such as generating full web pages, interactive visualizations, or even 2D animations from natural language.

1.1 Team

Kuzin Aleksey finding datasets, writing benchmarks evaluating pipeline.
Artyom Zaitsev finding datasets, accelerate evaluating via VLLM.
Kuzin Aleksey, Artyom Zaitsev computing and sharing resourses

2 Models Description

Table 1: Model Generation Configuration

Model Size	max gen. tokens	reasoning	precision	device	attnention
Qwen3-0.6B	1500	false	bfloat16	A100	Flash attention
Qwen3-8B	1500	false	bfloat16	A100	Flash attention
Qwen3-14B	1500	false	bfloat16	A100	Flash attention

Instead of reasoning mode, we inject "I have already thought" between <think> tokens. This approach was given from [Wenjie Ma, 2025]. The article shows that such an injection replaces the mode of reflection without much loss in accuracy. In our case, in this way we leave all tokens for a significant part of the generation - the svg code. We also used Qwen3-14B LoRA-finetuned, but dataset it was train on contained svg in only <path> tags without markers or labels. So as a result, the model has lost the ability to generate meaningful svg and we do not present its results.

To evaluate aesthetics(AScore furter) genereated svg's, we used pretrained model presented by [Schuhmann, 2023]. The regressor model that accepts clip embedding as input. The model was trained on a dataset, the aesthetics of which was evaluated as a weighted average of people's ratings from 0 to 10. So we convert svg to png before passing to the model. We have prepared a special dataset that contains descriptions and corresponding svg images. These images are very detailed and fit this description very well, so we considered the aesthetics of the condensed SVGs on this dataset. The dataset will be described later.

We can evaluate the beauty of the generated images, but it is also important for us to evaluate how well the image content matches the text query. To evaluate such a metric, we used the tifa benchmark [Hu et al., 2023]. How tifa benchmark does work? We have a prepared list of questions for each description. For each question, we have a valid list of answers. For example, we have an image description: "an empty room with a TV on the dresser." The question prepared by the benchmark is: "is there a TV in the room?", acceptable answers are yes/no. Next, we ask the LLM for this question for given description and filter the answer. Then we generate an svg image based on the specified description and feed it to the input of the VQA model. If the LM and VQA answers match, then we get a single score on this question, otherwise zero.

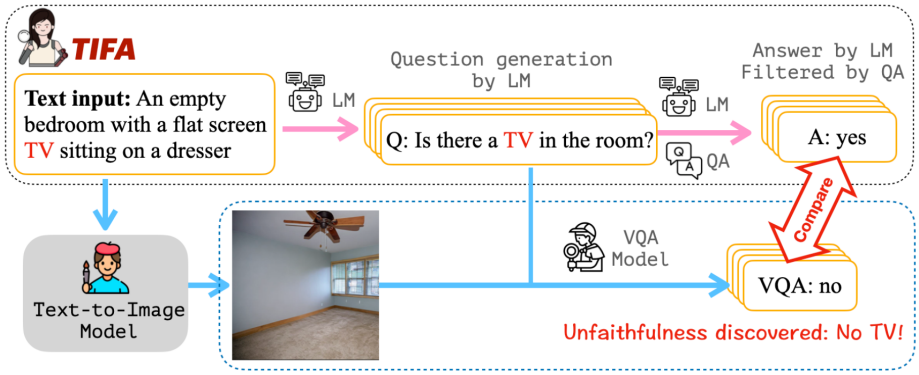

Figure 1: Overview.

3 Dataset

AScore evaluating dataset. [1]

	Valid
Descriptions	11188
Correct SVG	11170

Tifa benchmark dataset. [2]

	Valid
Descriptions	4097
Questions	25829

4 Evaluating

4.1 Metrics

$$\text{AvgAScore} = \frac{\sum_{i=1}^{n} \text{AScore}_i}{n} \tag{1}$$

$$\text{tifa}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1} \left(\text{LLM}(Q_i) = \text{VQA}(I, Q_i) \right), \tag{2}$$

where n - number of questions to j-s sample

$$\text{AvgTifa} = \frac{1}{m} \sum_{j=1}^{m} \text{tifa}_j, \tag{3}$$

[1] Is's scored by another model not interesting us. Donwload from kaggle this.
[2] Look tifa link above

где m - объём датасета

4.2 AScore

Model	AvgAScore	Percentage
Qwen3-0.6B	4.159800	98.030439
Qwen3-14B	4.263120	90.393912
Qwen3-8B	4.326931	83.133393

where is the percentage of correct generations in the last column.

4.3 Tifa

Table 2: Сравнение моделей по категориям

Category	Qwen/Qwen3-0.6B	Qwen/Qwen3-14B	Qwen/Qwen3-8B
activity	**0.231848**	0.223080	0.107682
animal/human	**0.195944**	0.195087	0.126821
attribute	0.276846	**0.277140**	0.172698
color	**0.480207**	0.473322	0.421113
counting	0.135903	0.137931	**0.165314**
food	**0.141603**	0.140505	0.120746
location	**0.152174**	0.151630	0.088587
material	**0.358852**	0.334928	0.272727
object	**0.156226**	0.155462	0.136618
other	**0.318408**	0.283582	0.313433
shape	0.550725	0.434783	**0.623188**
spatial	**0.377042**	0.377042	0.311258
Mean	**0.281315**	0.265374	0.238349

4.4 Visual assessment

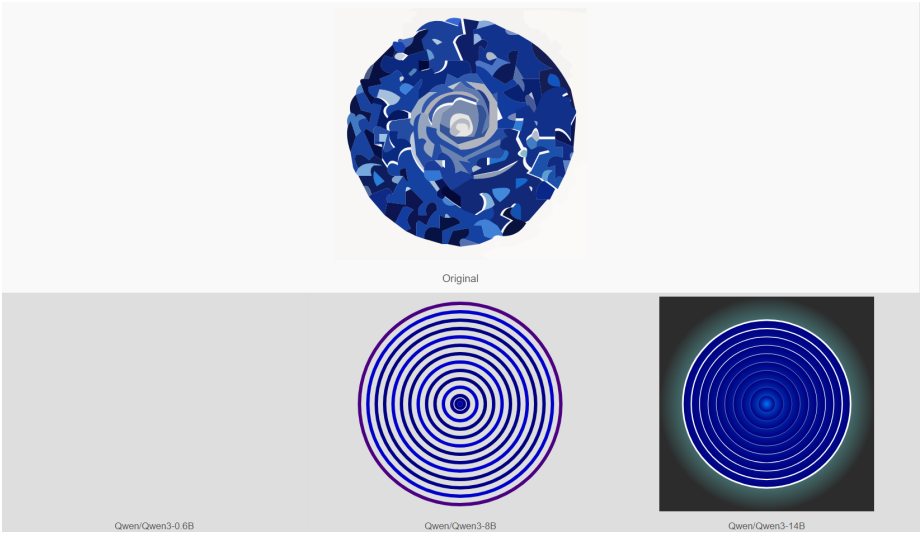

Figure 2: Example.

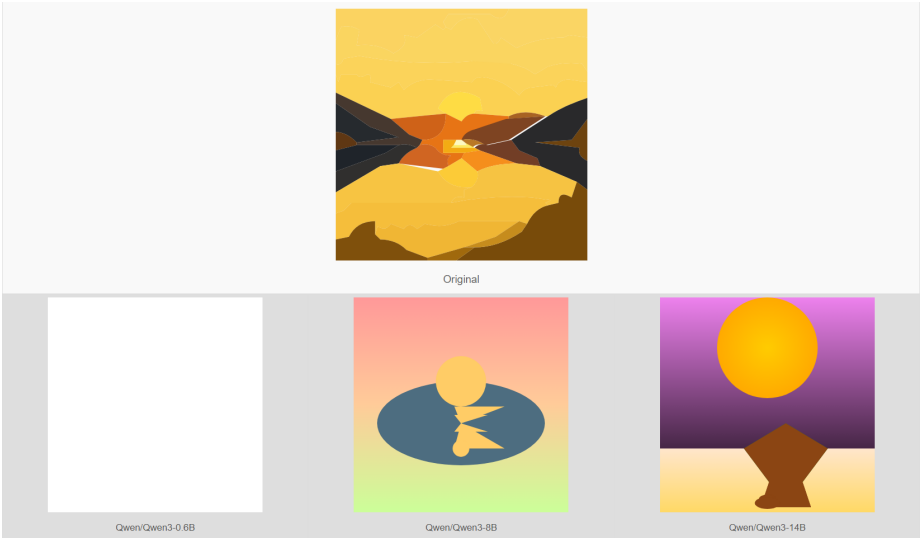

Figure 3: Example.

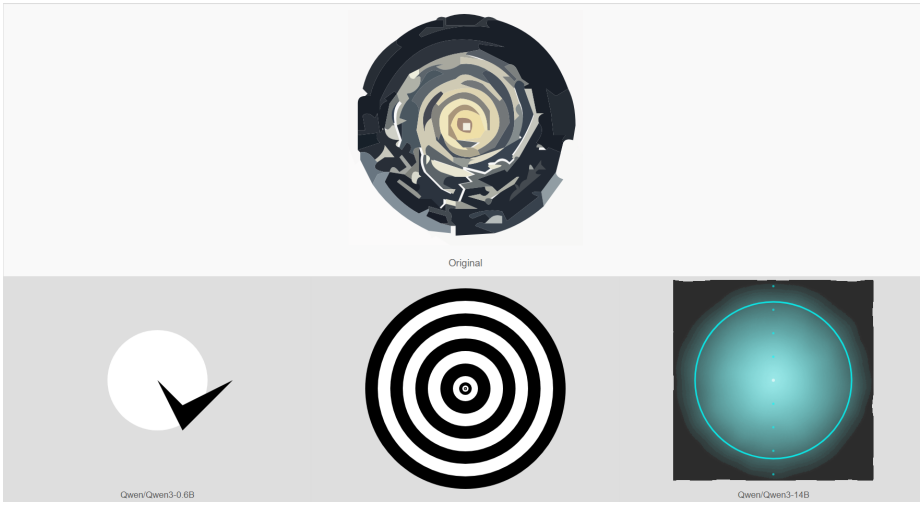

Figure 4: Example.

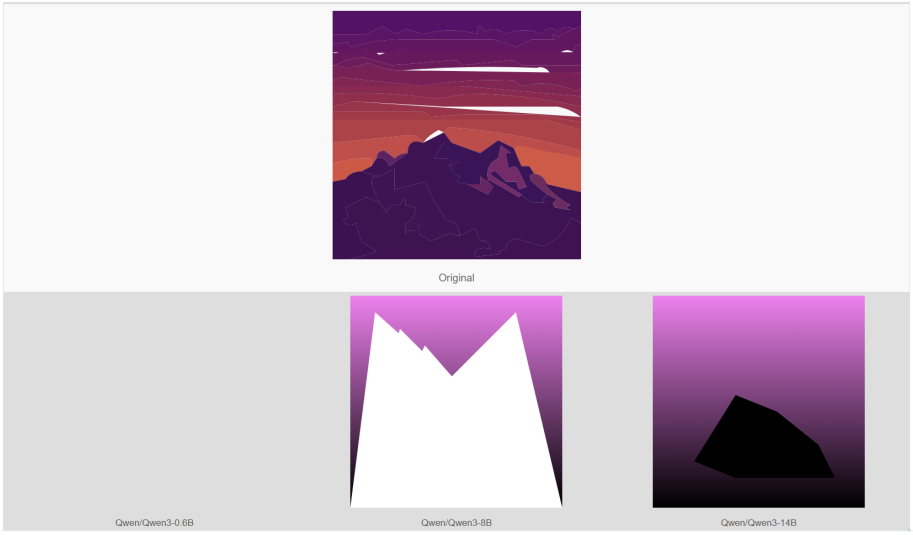

Figure 5: Example.

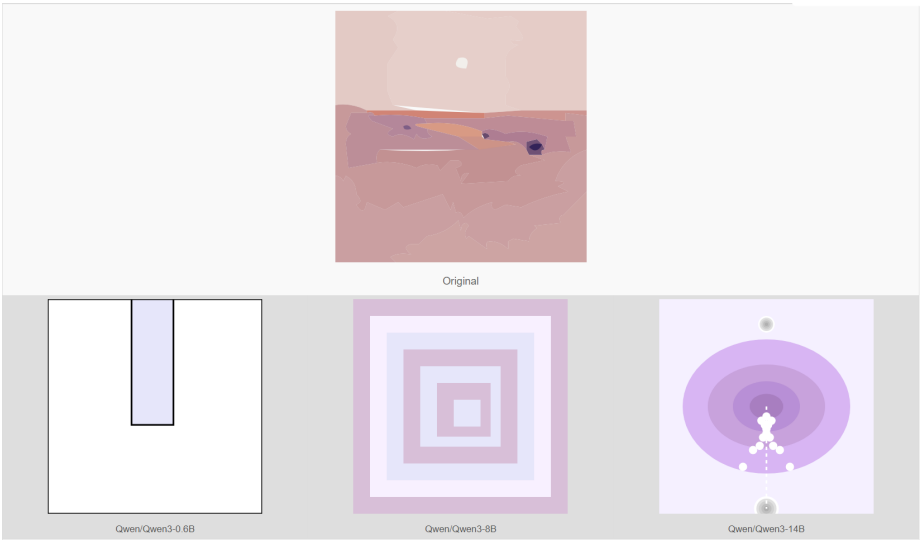

Figure 6: Example.

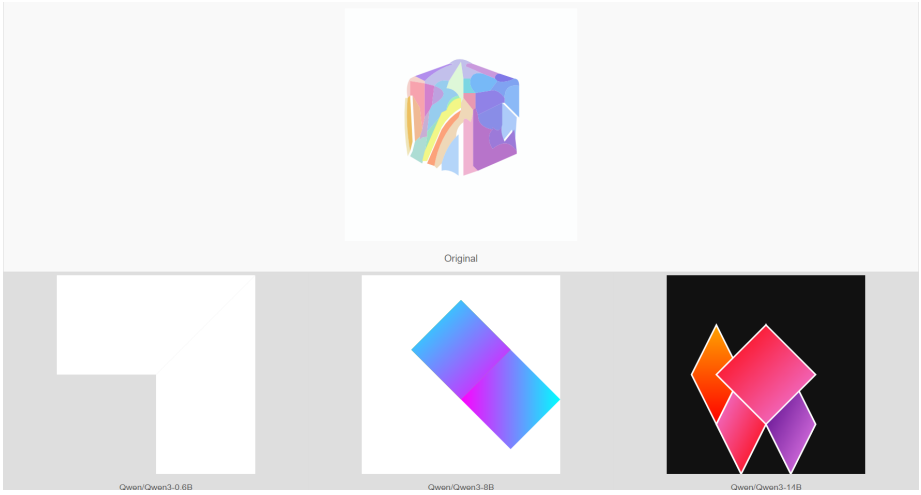

Figure 7: Example.

5 Conclusion

Benchmarks were calculated for three models, and a visual comparison was provided. The metrics of the tifa benchmark were surprising, the smallest model

showed the best result. Visually, looking not only at these examples, "Qwen-0.6B" drew as simple svg's as possible, and nevertheless got the best metrics. The aesthetics score is highest for the "Qwen-8B". Although adjusted for variance, it can be considered comparable to the largest Qwen.

References

[Hu et al., 2023] Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., and Smith, N. A. (2023). Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*.

[Schuhmann, 2023] Schuhmann, C. (2023). improved-aesthetic-predictor.

[Wenjie Ma, 2025] Wenjie Ma, Jingxuan He, C. S. T. G. S. M. M. Z. (2025). Reasoning models can be effective without thinking.