StableYolo: Optimizing Image Generation for Large Language Models

Harel Berger$^{3[0000-0001-6035-5127]}$, Aidan Dakhama$^{1[0009-0002-7318-7964]}$, Zishuo Ding$^{4[0000-0002-0803-5609]}$, Karine Even-Mendoza$^{1[0000-0002-3099-1189]}$, David Kelly$^{1[0000-0002-5368-6769]}$, Hector Menendez$^{1[0000-0002-6314-3725]}$, Rebecca Moussa$^{2[0000-0001-9123-6008]}$, and Federica Sarro$^{2[0000-0002-9146-442X]}$

[1] King's College London {aidan.dakhama,karine.even_mendoza,david.a.kelly, hector.menendez}@kcl.ac.uk
[2] University College London {r.moussa,f.sarro}@ucl.ac.uk
[3] Georgetown University bergerar0@gmail.com
[4] University of Waterloo zishuo.ding@uwaterloo.ca

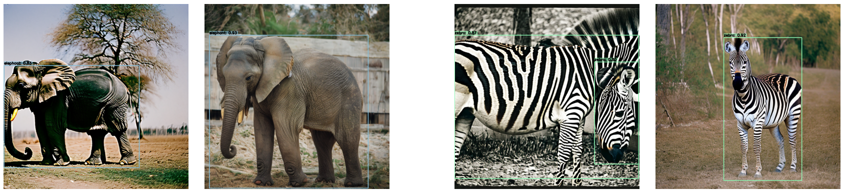
Abstract. AI-based image generation is bounded by system parameters and the way users define prompts. Both prompt engineering and AI tuning configuration are current open research challenges and they require a significant amount of manual effort to generate good quality images. We tackle this problem by applying evolutionary computation to Stable Diffusion, tuning both prompts and model parameters simultaneously. We guide our search process by using Yolo. Our experiments show that our system, dubbed StableYolo, significantly improves image quality (52% on average compared to the baseline), helps identify relevant words for prompts, reduces the number of GPU inference steps per image (from 100 to 45 on average), and keeps the length of the prompt short (≈ 7 keywords).

Keywords: LLMS · SBSE · Image Generation · Stable Diffusion · Yolo

1 Introduction

Generative AI [4] has proven to be one of the advances of the decade, disrupting the way AI contributes to Arts. Large Language Models (LLMs), in combination with generative models, provide interfaces that greatly scale up content generation. The training of these systems normally consists of a massive training dataset with associated descriptions. The correct description –or prompt– would create the desired outcome almost exactly. However, the level of vagueness and ambiguity in the description can lead to unexpected results or errors.

Image generation is a good example of this specific problem. When generating images, generative models like DALL-E, Midjourney or Stable Diffusion recommend variations to the prompt to guarantee that the quality of the generated image is closer to expectations [9]. For example, if the user aims to create a photorealistic image, it is recommended to include words like *"professional photoshoot, 8k resolution, photorealistic masterpiece"* or *"natural lighting"*. However,



(a) Elephant, before and after

(b) Zebra, before and after

Fig. 1: Before-and-after snapshots of optimization with StableYolo of two images: an elephant and a zebra.

it is not clear which words obtain the best results. Parameters of the algorithm, such as the number of steps used to generate an image, the initial random seed or the classifier-free guidance scale, significantly affect the outcome, as well. The goal is to create images under the specific condition of photo-realistic images.

To this end, we propose a framework, StableYolo, which aims to improve the quality of automatic image generation for photo-realistic images by leveraging the power of multi-objective search coupled with vision. The search process aims to improve different parameters of an image generation system (i.e., Stable Diffusion) according to the outcome of the chosen visualization system (i.e., Yolo). In this work, we use Stable Diffusion [2] and Yolo, respectively, but the framework can be extended to work with other systems. Our results show that StableYolo significantly improves the quality of the generated images and helps direct prompts effectively across various contexts (Fig. 1). StableYolo provides a 52% increase in quality compared to the baseline results.

Contributions The main contributions of this work are:
(1) Creating the first system combining search and vision for image generation;
(2) Testing this system on 42 different objects that can be obtained with image generation and detected by our vision system; and
(3) Obtaining significant improvements and prompt recommendations for the generation process of photo-realistic images.

2 Proposed Approach

Given a user-provided prompt of a photo-realistic image to a generative AI system, StableYolo attempts to improve image quality by finding an optimal combination of both, the prompt and the model's parameters. It uses a search-based process to identify an optimal combination, using Yolo's confidence estimates as the fitness function.

StableYolo uses a Genetic Algorithm (GA) for the search process, where each individual is a dictionary object with values taken from StableDiffusion's documentation [11]. These include: **(1)** *Number of iterations*: the number of iterations needed for the AI to go through the image (range [1,100]); **(2)** *Classifier-free guidance scale (CFG)*: a parameter that controls the prompt's influence on the resulting image (range [1,20]); **(3)** *Positive prompt*: enables the prompt to describe the images and their details; **(4)** *Negative prompt*: a sequence of keywords to be avoided during the image generation process; **(5)** *Seed*: the generation seed for randomization; and **(6)** *Guidance rescale*: prevents over-exposure by rescaling the guidance factor (range [0,1]). Positive prompt: In this work, we pass to the prompt, words that aim to produce more realistic images, according to different images extracted from the Photo-realistic Prompts documentation [10]; e.g., 'photograph', 'digital', 'color', 'Ultra Real' and 'award winning'. Negative prompt: We select references that might reduce the realistic component of the image; e.g., 'illustration', 'painting', 'drawing' and 'art'.

Utilizing GA, StableYolo optimizes both the LLM's parameters and the prompts based on the encoding of each potential individual. We start with a population of a set of individuals initialized with randomly selected parameters. Next, we evaluate the individuals using the fitness function. The fitter the individuals, the higher the chance they have of being chosen during the Tournament Selection process, after which the pair of individuals would undergo the crossover and mutation operators.

The crossover creates a pair of new offsprings by exchanging the attributes among the selected parents. After, the mutation operator (if applied) changes the parameters of an offpsring by uniformly assigning them new values within a given range. This results in the creation of a new population after each iteration of the algorithm. A stopping criterion is required to end the algorithm which is usually indicated by a number of generations specified by the user.

StableYolo aims to maximize its objective function through the following steps: (i) StableYolo creates positive and negative prompts and sets the parameters of Stable Diffusion to create four images per prompt; (ii) Yolo evaluates the images by identifying objects within it. These objects have an identification confidence that StableYolo uses to evaluate image quality; and (iii) StableYolo then calculates the fitness as the average quality of individual objects and images. This results in a different search objective based on the quality of each object. To avoid a multi-dimensional Pareto front, whose dimensions are unknown, the fitness calculates the object's quality average obtained with Yolo.

3 Experimental Setup and Results

In this section, we describe our research questions, experimental setup and results.

RQ We aim to answer the following research questions:

- **RQ1** To what extent can we improve the quality of an image generation system?
- **RQ2** What is the average prompt length to achieve maximum quality?
- **RQ3** What is the keyword frequency associated with quality images?

In order to provide a wide set of elements to evaluate StableYolo, we used 42 different categories of objects, animals and people that Yolo can recognize,

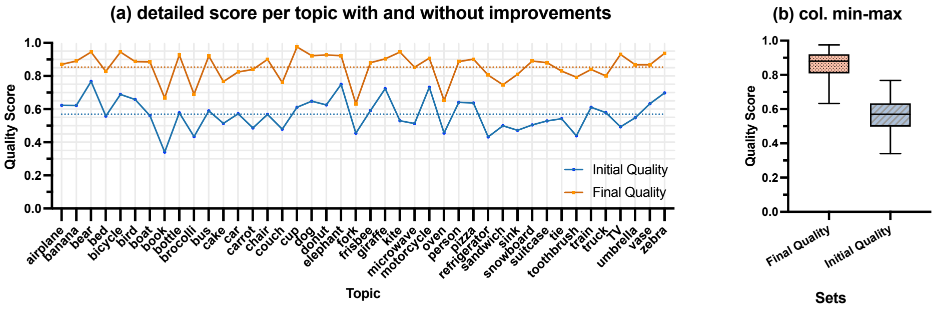


Fig. 2: **RQ1:** Quality Score Comparison: StableYolo (orange) vs. Non-Optimized (blue) Image Generation. StableYolo always produces better images on our dataset.

(e.g., person, elephant, zebra, dog, boat, and TV among others). For RQ2, we consider a $1-2$ word prompt as *small*, $3-6$ words as *medium* and 7+ words as *large*.

Experimental Setup For the GA, the settings were as follows; the population size was 25, the number of generations was set to 50, the mutation rate and crossover were both set to 0.2, and the tournament selection value was set to 5. Even given the small population size, we find the algorithm produces good results (Fig. 2a). The experiments were repeated 4 times. All experiments were conducted on a workstation running Ubuntu 20.04 LTS with 36 CPU cores, 256 GB RAM and an NVIDIA Titan V GPU with 12 GB memory.

Results Fig. 2 shows the results based on Yolo's quality assessment, ranked with a score from 0 to 1, where 1 is best. Fig. 2a shows the quality score with and without improvements, grouped by topic. The orange line represents the quality score per topic after applying StableYolo, while the blue line denotes the baseline before improvement (e.g., the topic baseline score is 0.558). The blue and orange dotted lines represent the mean quality score before and after improvement, respectively. Fig. 2b presents the quality score across all topics, where the orange boxplot shows the results with improvements, and the blue one shows them with no improvements. Fig. 2a shows improvement on all 42 prompts, with the orange line scores always being above the blue line scores. The initial prompt quality average is 0.567 with a standard deviation (SD) of 0.096, while the average score after optimization is 0.854 with an SD of 0.0836. This gives an average improvement of 52% in the quality scores after optimization. As the final quality scores fail the normality test, we applied the Wilcoxon test under the null hypothesis that both results follow the same distribution. The resulting p-value for the test is $4.55 \cdot 10^{-13}$, clearly rejecting the null hypothesis, and showing a statistically significant difference between both sets of results. The optimized set demonstrates a significant increase in the quality of the generated images.

Parameters and Prompts The average *optimal* value for the guidance scale is 9.65 ± 3.99 in the range [0,20] while the rescale value is 0.56 ± 0.29 on average. The

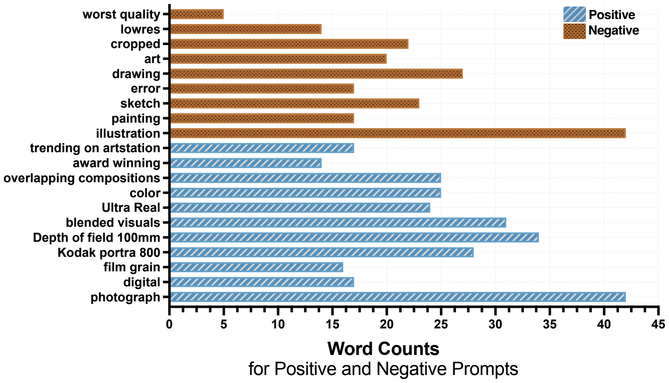


Fig. 3: **RQ3:** Positive and negative prompt terms counting for the best individuals.

number of inference steps to 45.3 ± 30.29 on average, in the range [0,100]. The positive prompt length on average is 6.5 ± 1.66 and the negative one is 4.45 ± 1.09. At the end of the optimization process, the most relevant keywords were found to be 'photograph' and 'Depth of field 100mm' for the positive prompt, and 'illustration' and 'drawing' for the negative one (Fig. 3). The memory cost to the GPU was 9 GB, with each prompt optimization taking approximately 2 hours to execute. On average, the GA takes 45.05 ± 10.01 iterations to converge. StableYolo can scale through multiple GPUs linearly for both the generation and the visualization process.

4 Related Work

Generative AI systems have been used to perform the task of image generation for many years [1,5,7,8]. Progress in this domain has resulted in incremental adjustments or entirely novel architectures, including but not limited to generator regularization [8], additional memory gate incorporation [12], use of dynamic thresholding [9], and fusion of neural architecture search with generative adversarial networks [6]. To the best of our knowledge, our work is the first to leverage the power of EAs to improve prompts and model parameters simultaneously. A comprehensive survey on other uses of LLMs can be found in Fan et al.'s work [3].

5 Conclusions

We introduce StableYolo, an innovative framework that enhances the fidelity of photo-realistic image generation by means of an iterative optimization procedure. StableYolo combines the text-to-image approach of Stable Diffusion v2 for image generation making use of the image-to-text mechanism of Yolo v8 for caption generation. An iterative process based on multi-objective search is adopted, utilizing affirmative and opposing prompts until convergence is attained. Opportunities for further enhancement persist, most notably the current iteration of StableYolo is constrained to classes identified by Yolo. Widening the spectrum of recognizable classes may enhance the diversity and analytical scope of generated images. Nonetheless, StableYolo exhibits a remarkable enhancement in the quality of image generation. In terms of generality, the StableYolo can easily be extended to other image generation models like DALL-E or Midjourney and different visualization systems apart from Yolo, for instance captioning systems, this will allow more objects and contexts to be addressed.

Availability StableYolo is at `https://github.com/SOLAR-group/StableYolo`.

Acknowledgments Authors are listed in alphabetical order. This work was supported by the ERC grant no. 741278 (EPIC), the UKRI TAS Hub grant no. EP/V00784X/1 and EP/V026801/2.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: IEEE/CVF CVPR. pp. 18208–18218 (2022)
2. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
3. Fan, A., Gokkaya, B., Harman, M., Lyubarskiy, M., Sengupta, S., Yoo, S., Zhang, J.M.: Large language models for software engineering: Survey and open problems (2023)
4. Jo, A.: The promise and peril of generative ai. Nature **614**(1), 214–216 (2023)
5. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: IEEE/CVF CVPR. pp. 10124–10134 (2023)
6. Li, W., Wen, S., Shi, K., Yang, Y., Huang, T.: Neural architecture search with a lightweight transformer for text-to-image synthesis. IEEE Transactions on Network Science and Engineering **9**(3), 1567–1576 (2022)
7. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: IEEE/CVF CVPR. pp. 12174–12182 (2019)
8. Mao, Q., Lee, H.Y., Tseng, H.Y., Ma, S., Yang, M.H.: Mode seeking generative adversarial networks for diverse image synthesis. In: IEEE/CVF CVPR. pp. 1429–1437 (2019)
9. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
10. Stable Diffusion Photorealistic Prompts: `https://prompthero.com/stable-diffusion-photorealistic-prompts`
11. The Stable Diffusion documentation: `https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/text2img`
12. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: IEEE/CVF CVPR. pp. 5802–5810 (2019)