High-Quality Images

GreenStableYolo: Optimizing Inference Time and Image Quality of Text-to-Image Generation

Jingzhi Gong (speaker)
University of Leeds, UK

Sisi Li
Beijing University of Posts and Telecommunications, China

Giordano d'Aloisio
Università degli Studi dell'Aquila, Italy

Zishuo Ding
University of Waterloo, Canada

Yulong Ye
University of Birmingham, UK

William B. Langdon
University College London, UK

Federica Sarro
University College London, UK

Background

Generative Artificial Intelligence (GenAI)

In recent years Generative AI has emerged as a powerful approach that enable machines to **generate different contents**, such as text, images, and videos.

Text-to-image generation

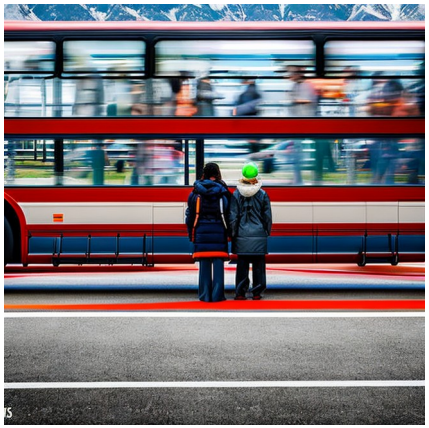
Particularly, text-to-image synthesis have garnered significant attention due to their potential in **translating human language and into visually meaningful representations**, facilitating tasks such as generating accompanying images for books, generating product images for advertising, and inspiring artists to create new forms of art.

Challenges

Achieving optimal image quality involves **tuning various parameters** such as # inference steps, positive and negative prompts.

Images generated with “two people and a

(a) Before parameter tuning: only one people, not clear



(b) After parameter tuning: clearly two people

StableYolo (SSBSE’24) shows that it is possible to use evolutionary algorithms to optimise image quality of Stable Diffusion

Challenges

GenAI models are **time and energy demanding**, largely contributing to increased CO2 emissions

"**1,000** image generation queries uses as much energy as **522** smartphone charges (11.49 kWh)"

— Power Hungry Processing: Watts Driving the Cost of AI Deployment?

	inference energy (kWh)	
task	**mean**	**std**
text classification	0.002	0.001
extractive QA	0.003	0.001
masked language modeling	0.003	0.001
token classification	0.004	0.002
image classification	0.007	0.001
object detection	0.038	0.02
text generation	0.047	0.03
summarization	0.049	0.01
image captioning	0.063	0.02
image generation	2.907	3.31

Table 2: Mean and standard deviation of energy per 1,000 queries for the ten tasks examined in our analysis.

Challenges

GenAI models are **time and energy demanding**, largely contributing to increased CO2 emissions

"**1,000** image generation queries uses as much energy as **522** smartphone charges (11.49 kWh)"

— Power Hungry Processing: Watts Driving the Cost of AI Deployment?

	inference energy (kWh)	
task	**mean**	**std**
text classification	0.002	0.001
extractive QA	0.003	0.001
masked language modeling	0.003	0.001
token classification	0.004	0.002
image classification	0.007	0.001
object detection	0.038	0.02
text generation	0.047	0.03
summarization	0.049	0.01
image captioning	0.063	0.02
image generation	2.907	3.31

Table 2: Mean and standard deviation of energy per 1,000 queries for the ten tasks examined in our analysis.

It is challenging to strike an optimal trade-off between
inference time vs. image quality

Advancing the State-of-the-Art: GreenStableYolo

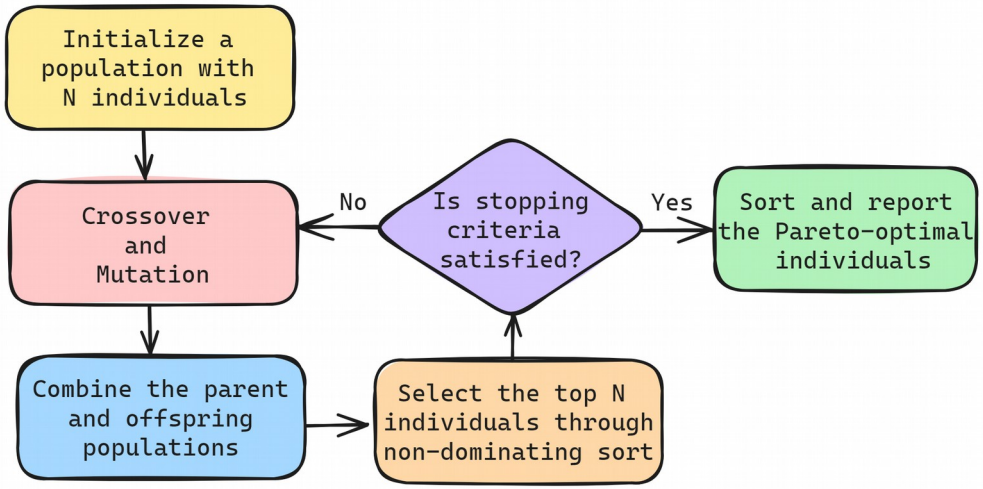
We propose a novel system, GreenStableYolo**,** that searches for an optimal trade-off between inference time and image quality by optimizing Stable Diffusion prompts and parameters

- Empirical evidence on GreenStableYolo **effectiveness** in achieving **significantly less inference** time and **higher hypervolume** compared to the state-of-the-art StableYolo.

- Analysis to understand the **influence of different parameters and prompts** on both inference time and image quality in Stable Diffusion.

Multi-Objective Evolutionary Algorithm

Non-dominated Sorting Genetic Algorithm (NSGA-II)



Fitness Functions:

(1)Inference time: the GPU time taken for the execution of the StableDiffusionPipeline function.

(2)Image quality: determined by performing object recognition with Yolo, then selecting objects that match the input prompt, and computing their average probabilities.

Settings

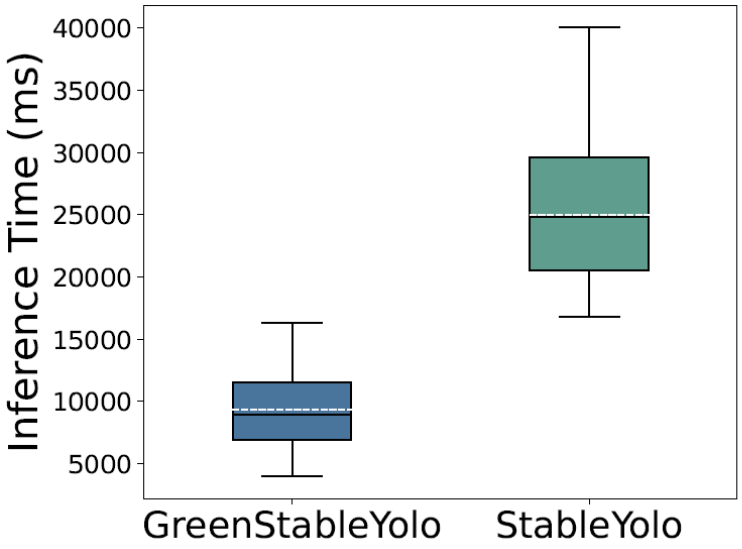
(a) No negative prompts: the elephant has two noses

(b) With negative prompts like “illustration, painting, art”: much more realistic

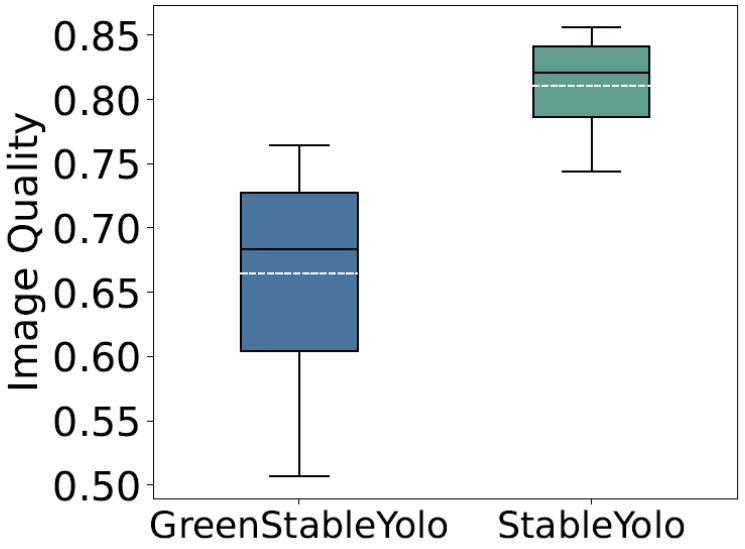
Figure: Two images generated with “an elephant wearing a party hat”.

- To make a straightforward comparison with StableYolo, we adopt **the settings** used in previous work, e.g., the following **parameters** are tuned:
  (1) Inference steps (1 to 100);
  (2) Guidance scale (1 to 20);
  (3) Guidance rescale (0 to 1);
  (4) Randomized Seed (1 to 512);
  (5) Positive prompt, e.g., “photograph”, “color”, and “ultra real”; and
  (6) Negative prompt, e.g., “sketch”, “cropped”, and “low quality”.
- We selected **weights** of 0.001 for image quality and -1000 for inference time based on empirical investigation of different weight combinations.
- To assess variability, we evaluated each model **15 times** using different random seeds, focusing solely on the prompt **“two people and a bus”** due to time constraints.

RQ1: To what extent can GreenStableYolo improve image quality and inference time compared with StableYolo?



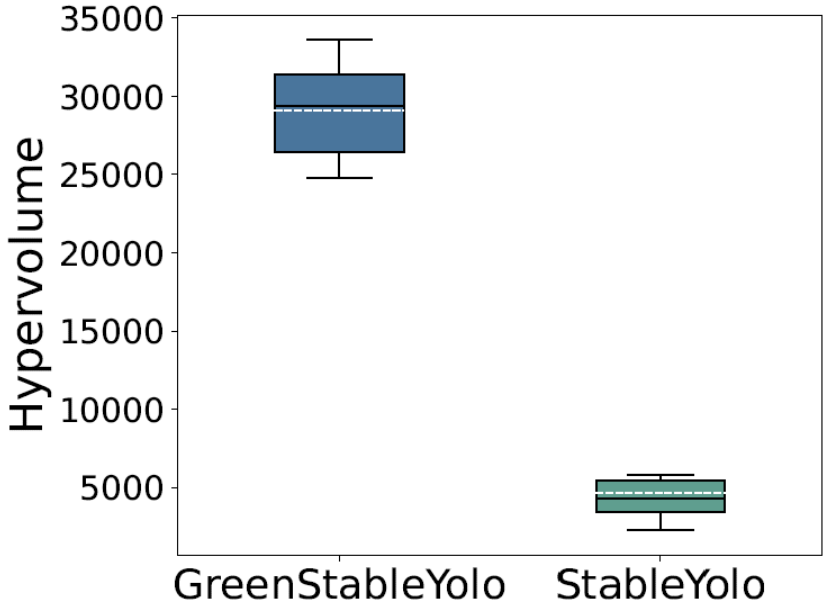
(a) Inference time (ms)



(b) Image quality

- Figure (a) reveals that GreenStableYolo achieves an average inference time of 9.4 seconds. Conversely, StableYolo exhibits an average inference time of 25 seconds, **which is 1.66 times slower**.
- However, in Figure (b), GreenStableYolo experiences approximately an **average degradation of 0.18 times in image quality**.

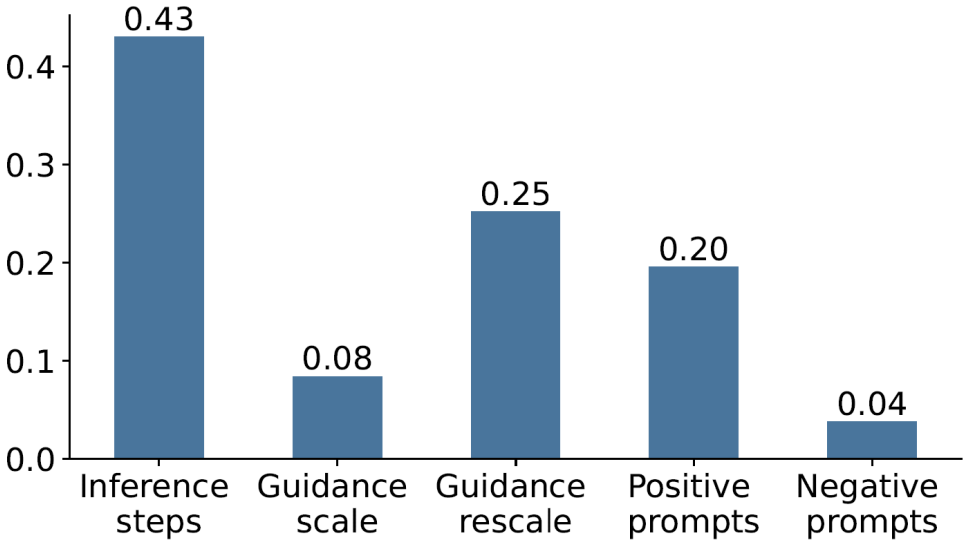
RQ1: To what extent can GreenStableYolo improve image quality and inference time compared with StableYolo?



(c)
Hypervolume

- Figure (c) shows GreenStableYolo achieves an average hypervolume of 29074, surpassing StableYolo's score of 4642 by **5.26 times**.
- This substantial difference demonstrates the clear **dominance of GreenStableYolo** over StableYolo in this two objective optimization problem for text-to-image generation.

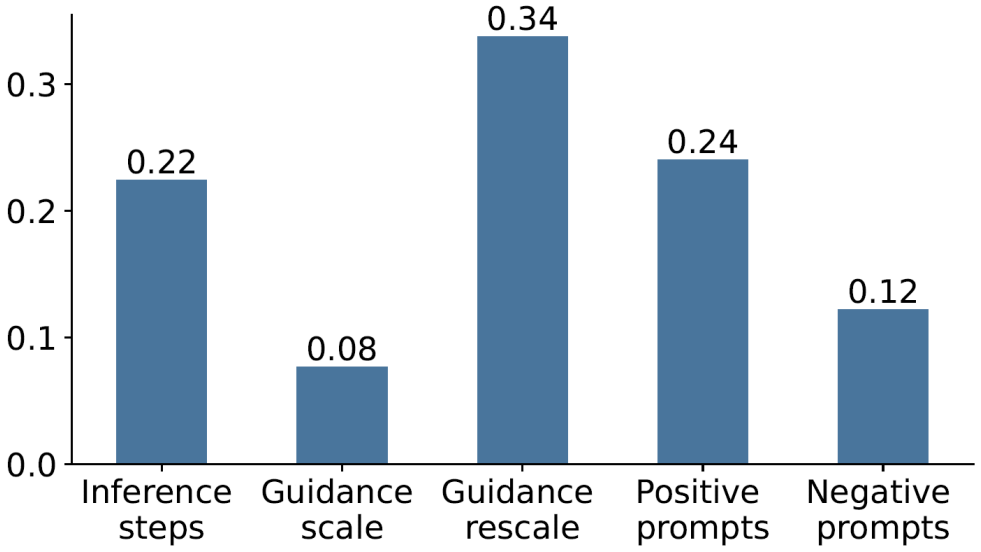
RQ2: How do parameters of Stable Diffusion influence the inference time for image generation?



(d) Importance w.r.t. inference time

- We followed an ICSE work and built two Random Forest regressors to compute the **Gini importance** of each parameter.
- We found that the **number of inference steps** emerges as a significant factor affecting inference time.
- This is expected, as more steps involve **more computations**, thereby resulting in higher inference time.

RQ3: How do parameters of Stable Diffusion influence the image quality for image generation?



(e) Importance w.r.t. image quality

- As for the image quality, parameters like **guidance rescale** and **positive prompts** play a relatively more critical role, which confirms the value of our work:
  - Simply increasing computational resources do not necessarily translate to better image quality;
  - Instead, identifying optimal parameters to **balance computational efficiency and output quality** are more crucial.

Conclusion and Future Directions

- In text-to-image generation, achieving images of **high-quality** is often not the only important aspect to consider, as **inference time**, which directly impacts user experience and energy consumption, also plays a critical role.

- In this work we introduced GreenStableYolo, a **pioneering eco-friendly** approach leveraging NSGA-II to strike an **optimal trade-off** between these two objectives for Stable Diffusion.

- **Future research** can expand upon our evaluation by incorporating alternative initial prompts, optimizing different performance metrics such as energy consumption, and broadening to other Generative AI systems such as DALL-E, ImageFX, or Midjourney.

- Repository available at https://github.com/gjz78910/GreenStableYolo.

Image generated with