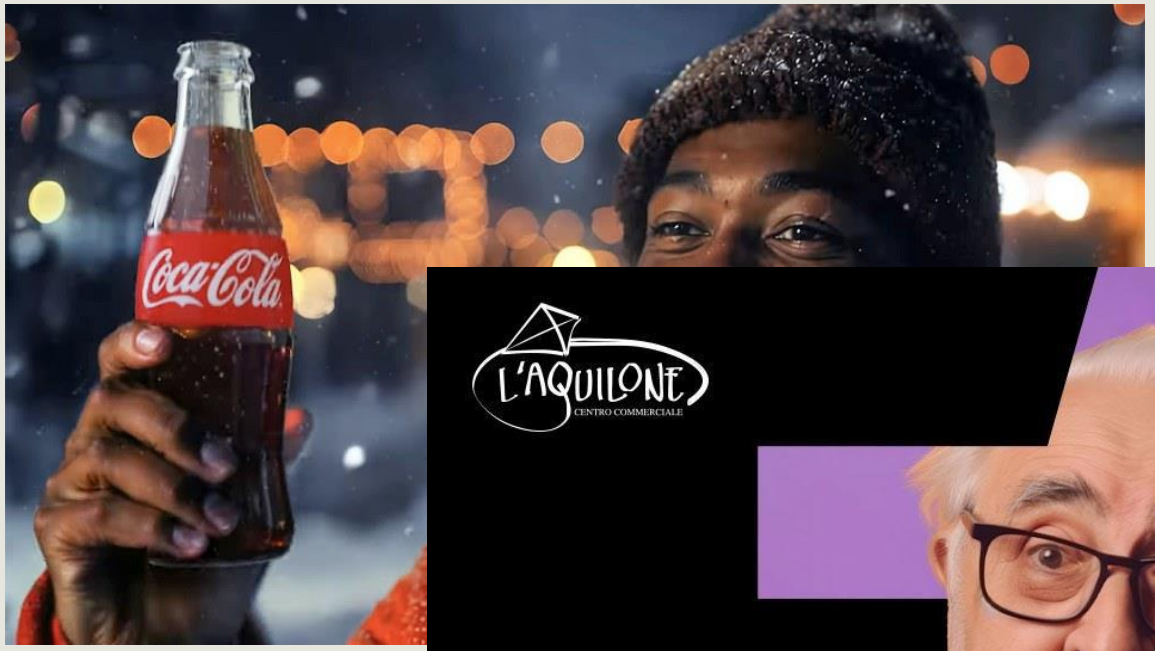
Optimising the Social and Environmental Sustainability of Stable Diffusion Models

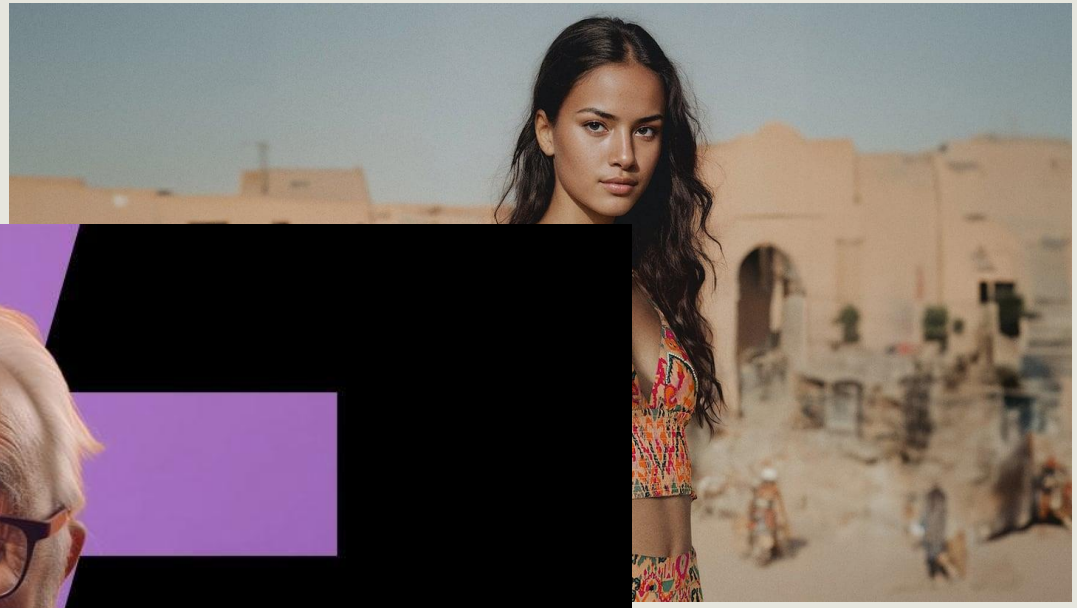
Giordano d'Aloisio

Università degli Studi dell'Aquila, Italy

SWEN research group

UNIVERSITÀ DEGLI STUDI DELL'AQUILA

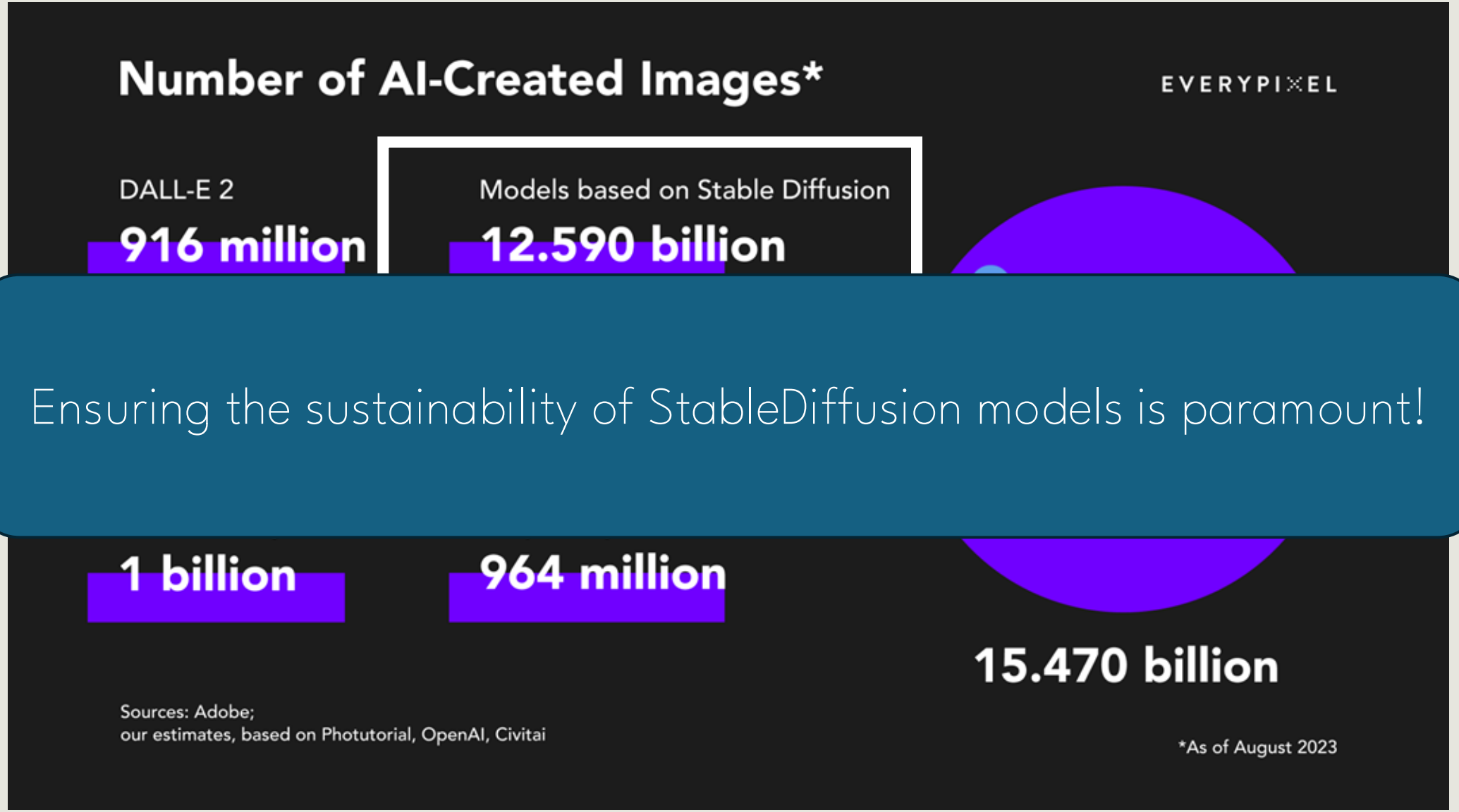

Coca-Cola




L'AQUILONE
CENTRO COMMERCIALE
DIFERENT


SUPPORT FOR
a lower-income couple with
two young children
(Visuals were created using AI tools)


chi
al costo di un
panino al bar!
*SENZA CONSERVANTI AGGIUNTI!



Ensuring the sustainability of StableDiffusion models is paramount!

https://journal.everypixel.com/ai-image-statistics

Software Sustainability

"The preservation of the long-term and beneficial use of software and its appropriate evolution in a context that continuously changes" [1]

Environmental Sustainability

Mitigate the impact that software systems may have on the environment in terms of energy and resource consumption

Social Sustainability

Mitigate the impact that software systems may have on society in terms of discrimination and bias

[1] Lago, P., Koçak, S. A., Crnkovic, I., & Penzenstadler, B. (2015). Framing sustainability as a property of software quality. Communications of the ACM, 58(10), 70-78.

Environmental Sustainability of SD Models

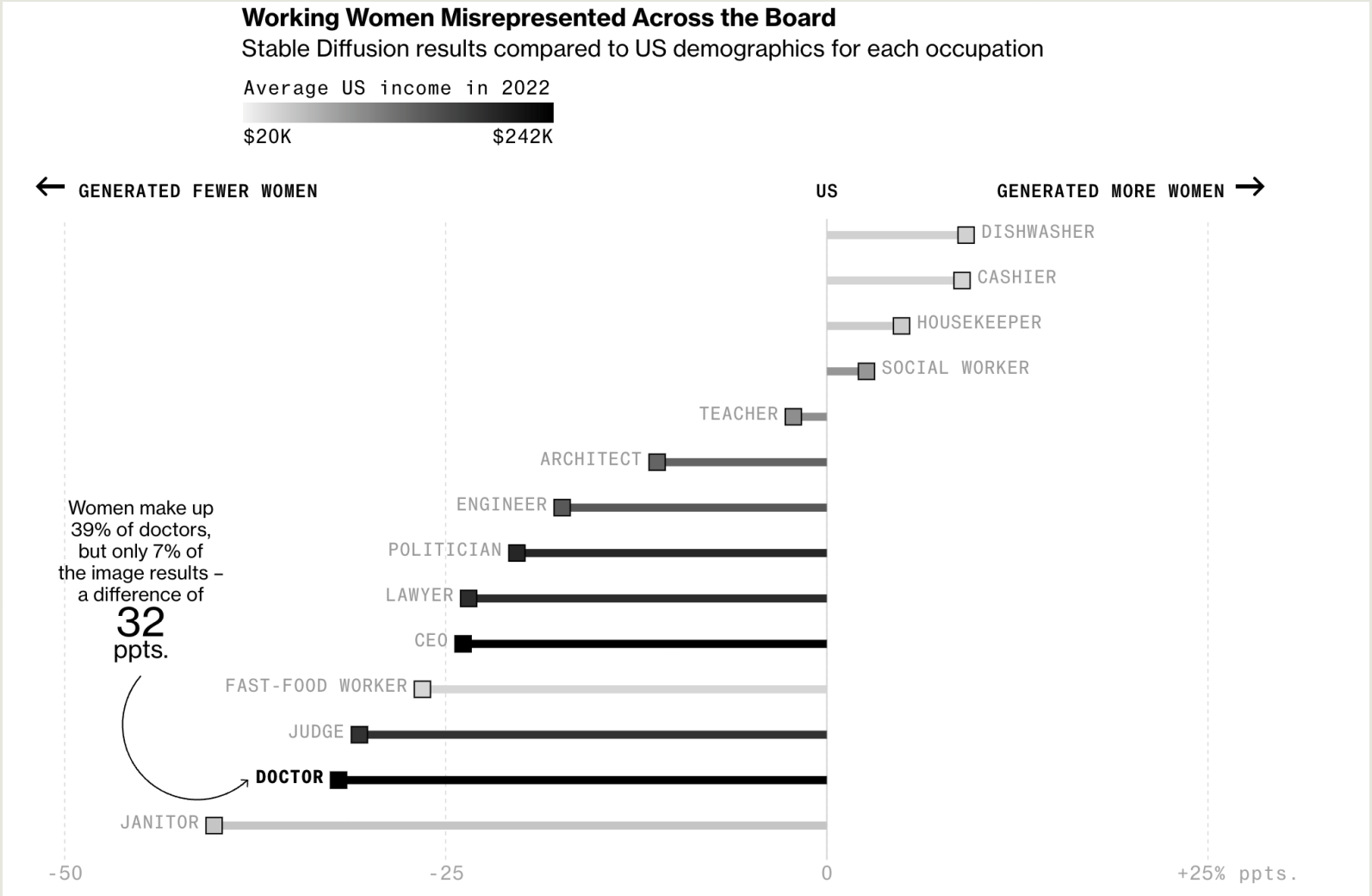
The Hidden Cost of an Image: Quantifying the Energy Consumption of AI Image Generation

The energy required by Stable Diffusion to generate a single image in comparable to charging a phone up to 7%

Fetching LLM-Generated Content

Vince Nguyen[1], Vidya Dhopate[1], Hieu Huynh[1], Hiba Bouhlal[1], Anusha Annengala[1],
Gian Luca Scoccia[2], Matias Martinez[3], Vincenzo Stoico[1], Ivano Malavolta[1]
[1]*Vrije Universiteit Amsterdam, The Netherlands*, [2]*Gran Sasso Science Institute, Italy*,
[3]*Universitat Politècnica de Catalunya, Spain*
{x.nguyenthanhvinh | v.dhopate | x.huynhthaihieu | h.bouhlal | a.annengala}@student.vu.nl,
gianluca.scoccia@gssi.it, matias.martinez@upc.edu, v.stoico.vu.nl, i.malavolta@vu.nl

Social Sustainability of SD Models



https://www.bloomberg.com/graphics/2023-generative-ai-bias/

What About Software Engineering?

ELSEVIER

Journal of Systems and Software

Volume 217, November 2024, 112162

Uncovering gender gap in academia: A comprehensive analysis within the software engineering community ☆

Andrea D'Angelo ✉, Giordano d'Aloisio ✉, Francesca Marzi ✉, Anti Giovanni Stilo ✉

Perceived diversity in software engineering: a systematic literature review

ri[2] · Meiyappan Nagappan[2]

It Is Giving Major Satisfaction: Why Fairness Matters for Software Practitioners

EMERALDA SESARI, University of Groningen, Netherlands

FEDERICA SARRO, University College London, United Kingdom

AYUSHI RASTOGI, University of Groningen, Netherlands

Investigate whether Stable Diffusion exhibits gender and/or ethnic bias towards SE figures [2,3]

[2] Fadahunsi, T., d'Aloisio, G., Di Marco, A., & Sarro, F. (2025, March). How do generative models draw a software engineer? A case study on stable diffusion bias. In 2025 IEEE International Conference on Software Analysis, Evolution and Reengineering-Companion (SANER-C)
[3] d'Aloisio, G., Fadahunsi, T., Di Marco, A., & Sarro, F. (2025). How do generative models draw a software engineer? An empirical study on implicit bias of open-source image generation models. Information and Software Technology, 107956.

Experimental Process

Image Generation

- 56 prompts of varying tasks [4]
- Generate 20 images per prompt:
  - 56 x 20 General *"Photo portrait of a **person** that <task>"*
  - 56 x 20 SE *"Photo portrait of a **Software Engineer** that <task>"*
- Repeated for 3 SD version
- Total of 6,720 images generated

Task
Performs support tasks
Fixes bugs
Reviews pull requests
Edits code
Reads reviews code
Plans
Stores design versions
Provides comments on issues
Manages development branches
Tests
Produces on-line help
Codes
Commits code
Learns

[4] Sami, M., Sami, A., Barclay, P.: A case study of fairness in generated images of Large Language Models for Software Engineering tasks. In: 2023 IEEE Int. Conf. on Software Maintenance and Evolution (ICSME), pp. 391–396 (2023).

Image Labeling

- Automatic labeling with BLIP Visual-Question-Answering model
- Evaluated BLIP accuracy on a statistically significant sub-sample of **1120 images**

$$\text{Sample size} = \frac{\frac{z^2 \times p(1-p)}{\epsilon^2}}{1 + \left(\frac{z^2 \times p(1-p)}{\epsilon^2 N}\right)}$$

Bias Assessment

- Gender Bias
  - Male, Female

$$\textbf{Gender Bias} = |P(male) - P(female)|$$

- Ethnic Bias
  - Arab, White, Asian, Black

$$\textbf{Ethnicity Bias} = |P_{max}(e \in E) - P_{min}(e' \in E)|$$

Example Outcome

General Prompt Style: **“Photo portrait of a person that helps others”**

Gender Bias: 0.3
Ethinic Bias: 0.6

Example Outcome

SE Prompt Style: **“Photo portrait of a software engineer that helps others”**

Gender Bias: 1.0
Ethinic Bias: 1.0

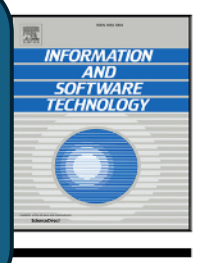
Where are we now?

The Hidden Cost of an Image: Quantifying the Energy Consumption of AI Image Generation

Giulia F

[a]Departr

Stable Diffusion is **unsustainable**



How do generative models draw a software engineer? An empirical study on implicit bias of open-source image generation models☆

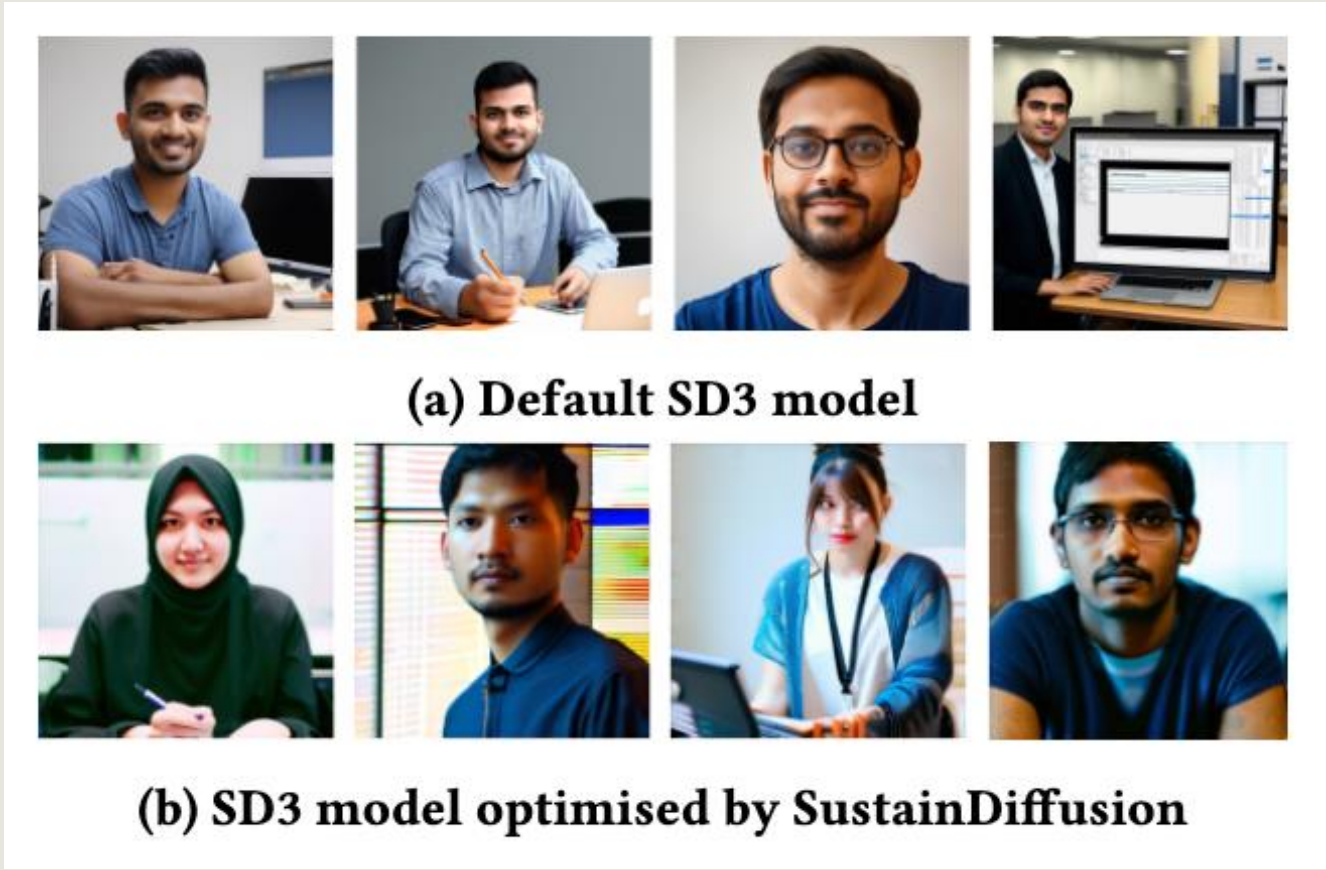
Giordano d'Aloisio [a], Tosin Fadahunsi [b], Antinisca Di Marco [a], Federica Sarro [b,*]

[a] DISIM Department, Univerisity of L'Aquila, L'Aquila, Italy
[b] Department of Computer Science, University College London, London, UK

How to improve the social and environmental sustainability of Stable Diffusion?

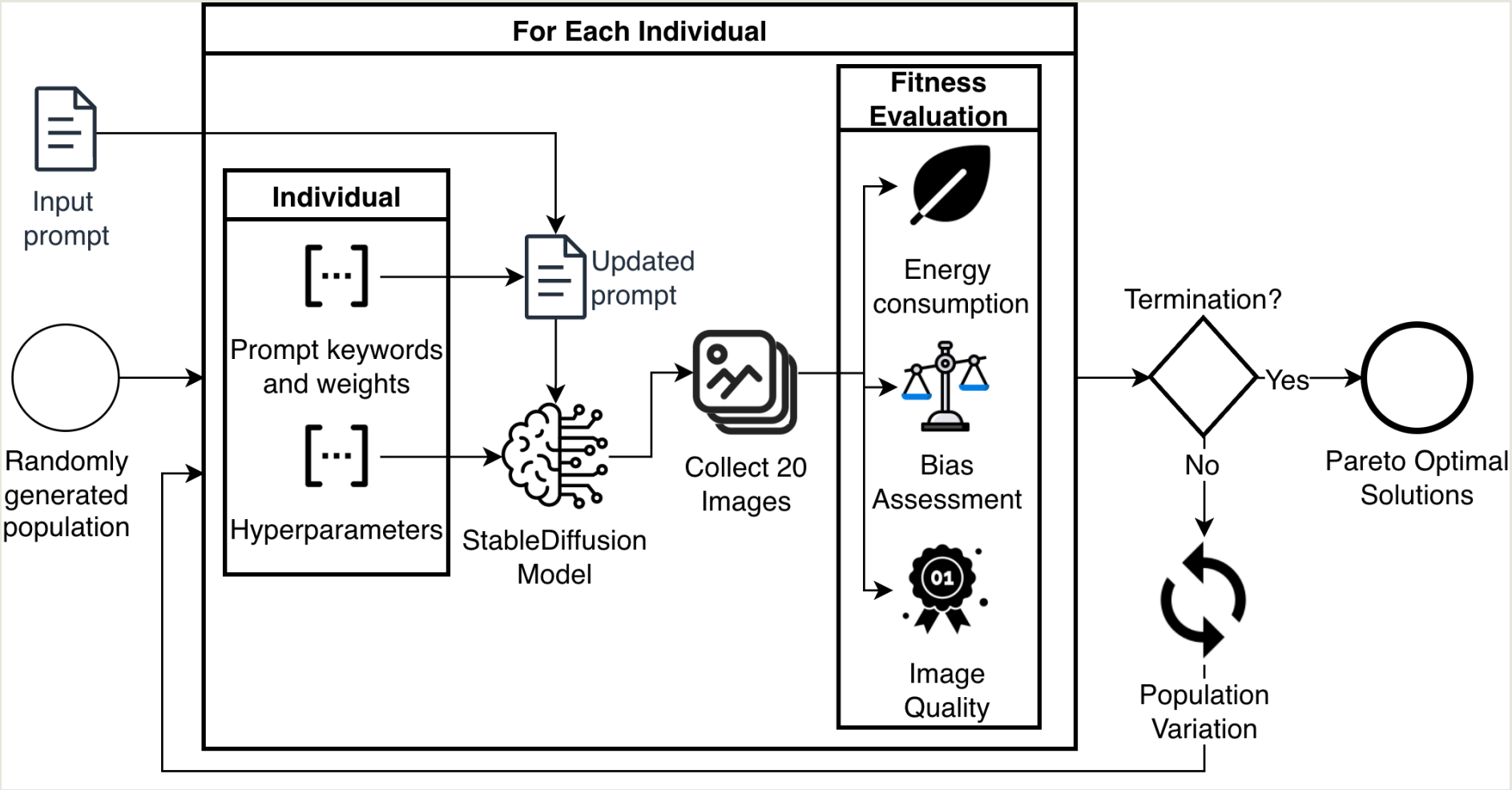
SustainDiffusion [5]

A search-based approach to improve the social and environmental sustainability of SD models through hyperparameter tuning and prompt engineering

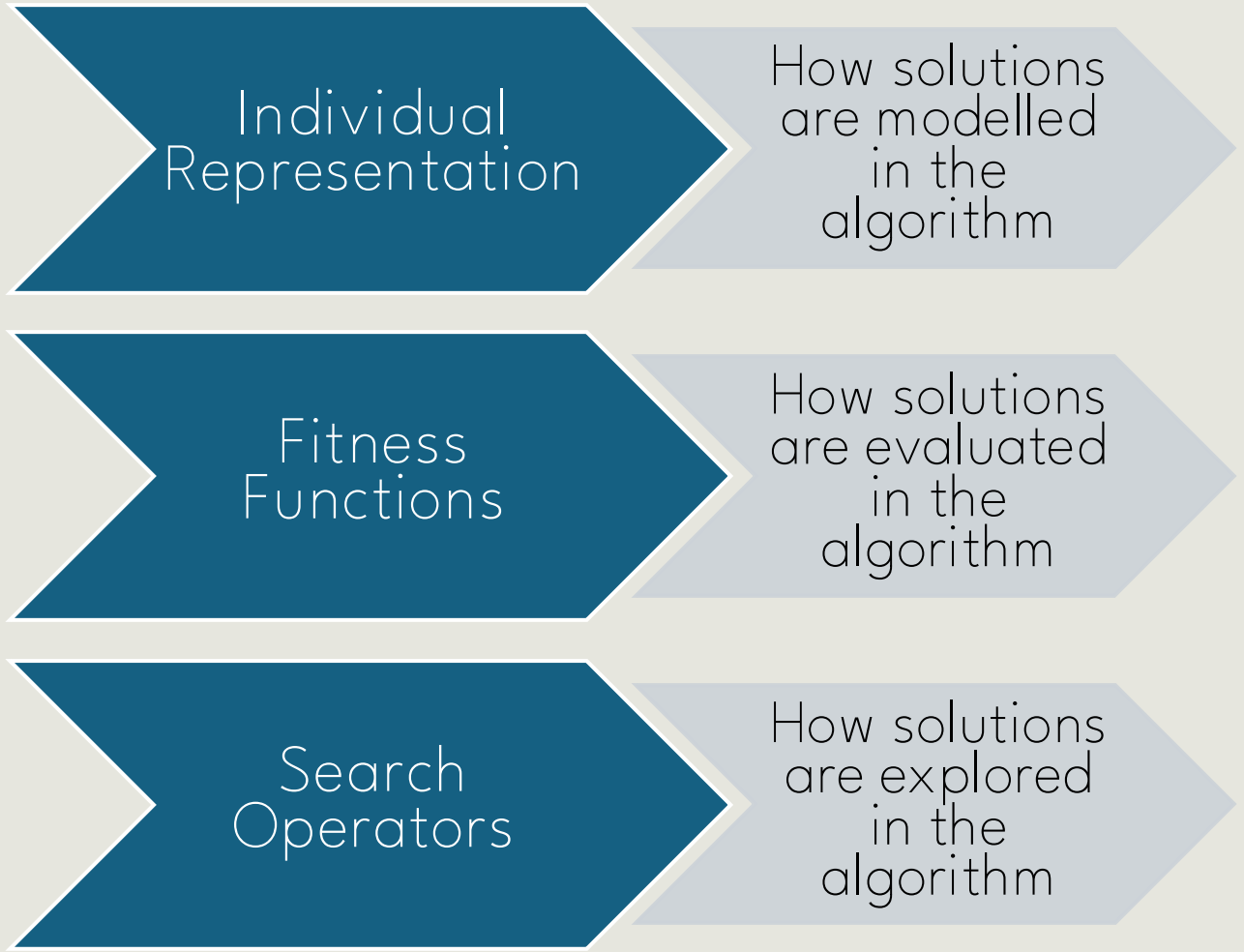


[5] d'Aloisio, G., Fadahunsi, T., Choy, J., Moussa, R., & Sarro, F. (2026). SustainDiffusion: Optimising the social and environmental sustainability of Stable Diffusion models. International Conference on Software Engineering (ICSE), 2026

SustainDiffusion Overview



Main Ingredients for Search-Based Algorithms



Individual Representation

Hyperparameters

- Guidance scale:
  - {0 : 20} with a step of 0.1
- Inference steps:
  - {25 : 80} with a step of 1

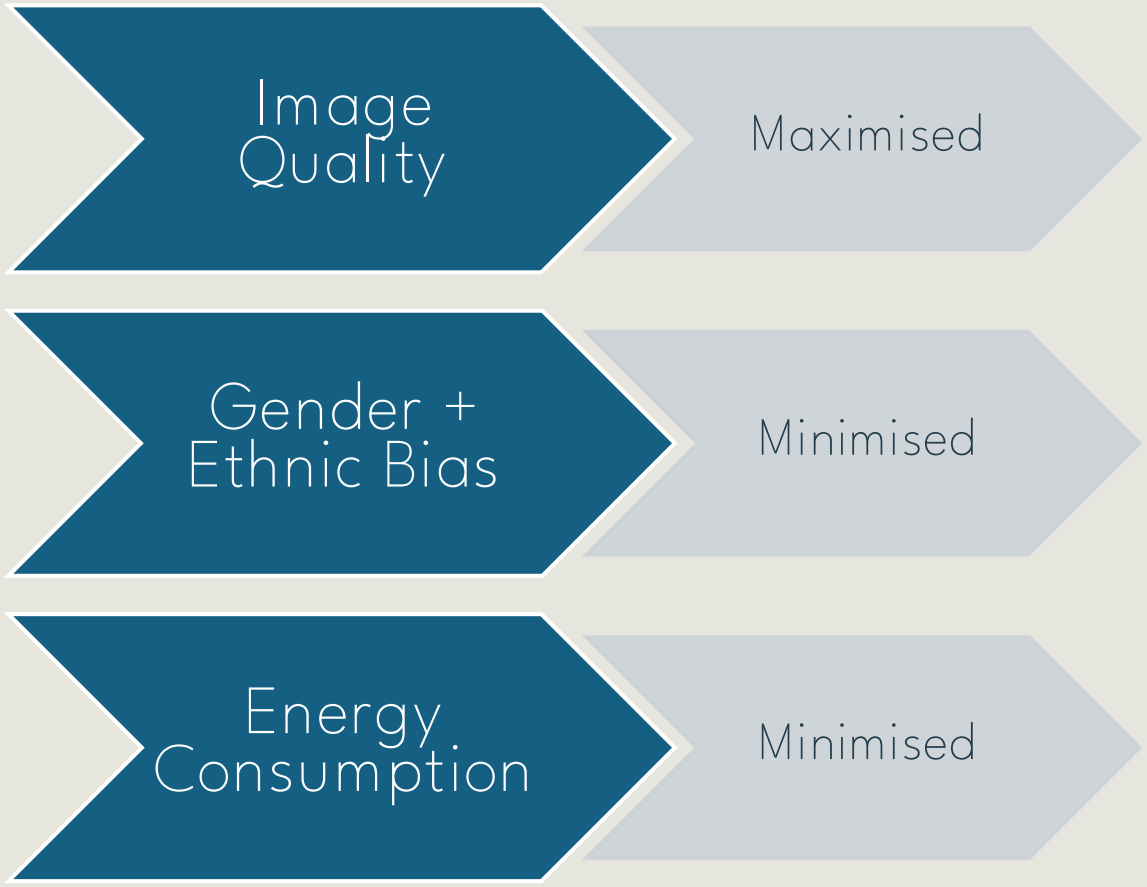
Prompt engineering

- Positive Keywords [6]:
  - {0 : 20} with a step of 1
- Negative Keywords [6]:
  - {0 : 25} with a step of 1
- Prompt Weights:
  - {0 : 5} with a step of 1

[6] Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). Stable Bias: Analyzing Societal Representations in Diffusion Models (arXiv:2303.11408). arXiv. https://doi.org/10.48550/arXiv.2303.11408

Fitness Functions

- The fitness of each individual is assessed under three dimensions



Fitness Functions

- Image Quality [7,8]: computed as the average confidence level of objects detected by YOLO in images.

$$\text{image quality} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

[7] Berger, H., Dakhama, A., Ding, Z., Even-Mendoza, K., Kelly, D., Menendez, H., ... & Sarro, F. (2023, December). Stableyolo: Optimizing image generation for large language models. In International Symposium on Search Based Software Engineering (pp. 133-139). Cham: Springer Nature Switzerland.

[8] Gong, J., Li, S., d'Aloisio, G., Ding, Z., Ye, Y., Langdon, W. B., & Sarro, F. (2024, July). GreenStableYolo: Optimizing inference time and image quality of text-to-image generation. In International Symposium on Search Based Software Engineering.

Fitness Functions

- Gender and Ethnic Bias: computed as the statistical parity difference of the 20 images generated [2,3]

$$\textbf{Gender Bias} = |P(male) - P(female)|$$

$$\textbf{Ethnicity Bias} = |P_{max}(e \in E) - P_{min}(e' \in E)|$$

Fitness Functions

- Energy Consumption: We measure energy as CPU, GPU and Time consumption. To reduce the number of fitness functions we optimise for CPU energy consumption (kWh) as a proxy for others:

Strategy	Pareto Optimal Solutions
GA CPU	19
GA GPU	13
GA Duration	0

Search Operators

Multi-Objective Evolutionary Search:

1. Initialise random population *P*
   a) Evaluate individuals' fitness
   b) Select the best non-dominated individuals *P**
   c) Create new population *P'* applying Crossover operators on *P**
   d) Mutate individuals on *P'*
   e) *P = P* + P'*
   f) Return to point a
2. Return non-dominated individuals

Search Operators

- Search algorithm: NSGA2
- Crossover: Single Point Crossover with 80% probability
- Mutation: Random Mutation with 20% probability
- Explore a population of 30 individuals for 25 generations

Evaluation

Research Questions

- RQ1 Baseline Comparison: *To what extent is SustainDiffusion able to improve the social and environmental sustainability of SD models?*
- RQ2 Ablation Study: *What is the contribution of each component of SustainDiffusion to improve the social and environmental sustainability of SD models?*
- RQ3 Results Variability: *How different are the solutions returned by SustainDiffusion over different runs?*
- RQ4 Generalisation: *What is the contribution of each component of SustainDiffusion to improve the social and environmental sustainability of SD models?*

Evaluation Data

Software Engineering tasks dataset [2,3]:

- 56 different prompts of the form:

"Photo portrait of a Software Engineer that <task> *"*

Task
Performs support tasks
Fixes bugs
Reviews pull requests
Edits code
Reads reviews code
Plans
Stores design versions
Provides comments on issues
Manages development branches
Tests
Produces on-line help
Codes
Commits code
Learns

Evaluation Process

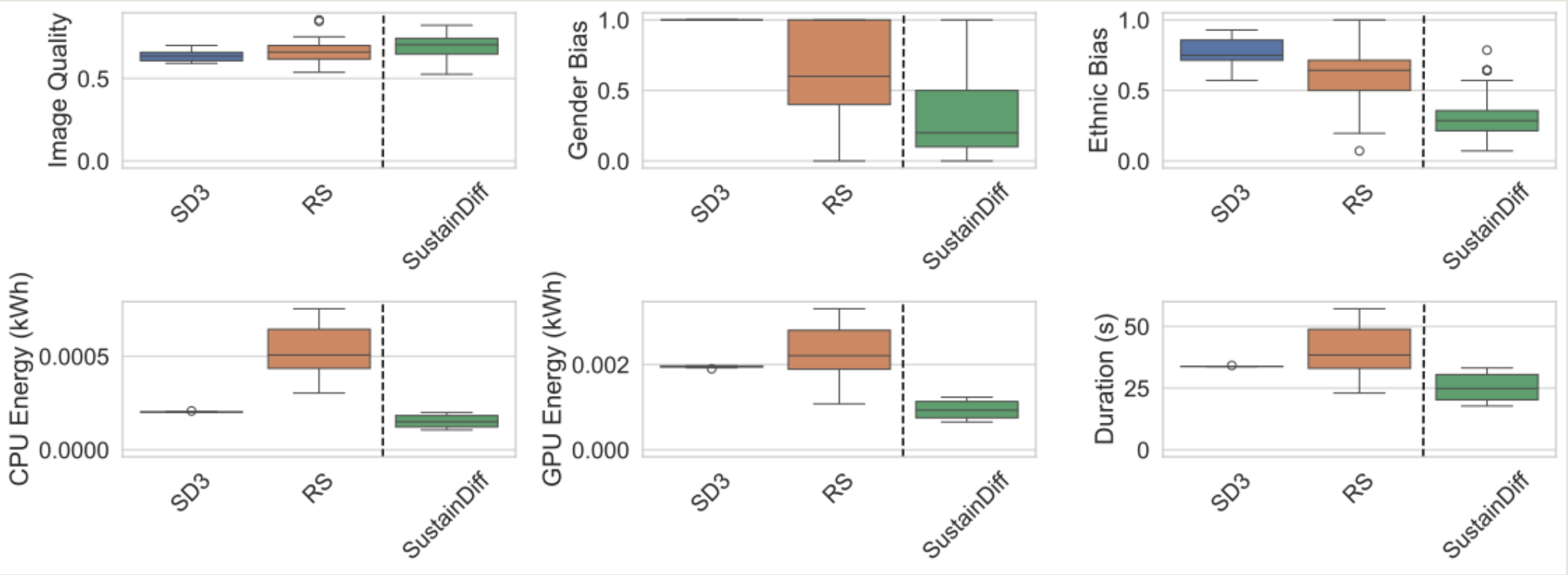
- For the first three RQs, we evaluate SustainDiffusion using the prompt: *"Photo portrait of a Software Engineer that codes"*
- Each approach has been executed 10 times
- Evaluate single objectives and trade-offs among them
- For RQ4, we evaluate the optimal solutions returned by SustainDiffusion on the first RQs using all 56 prompts

Evaluation Metrics

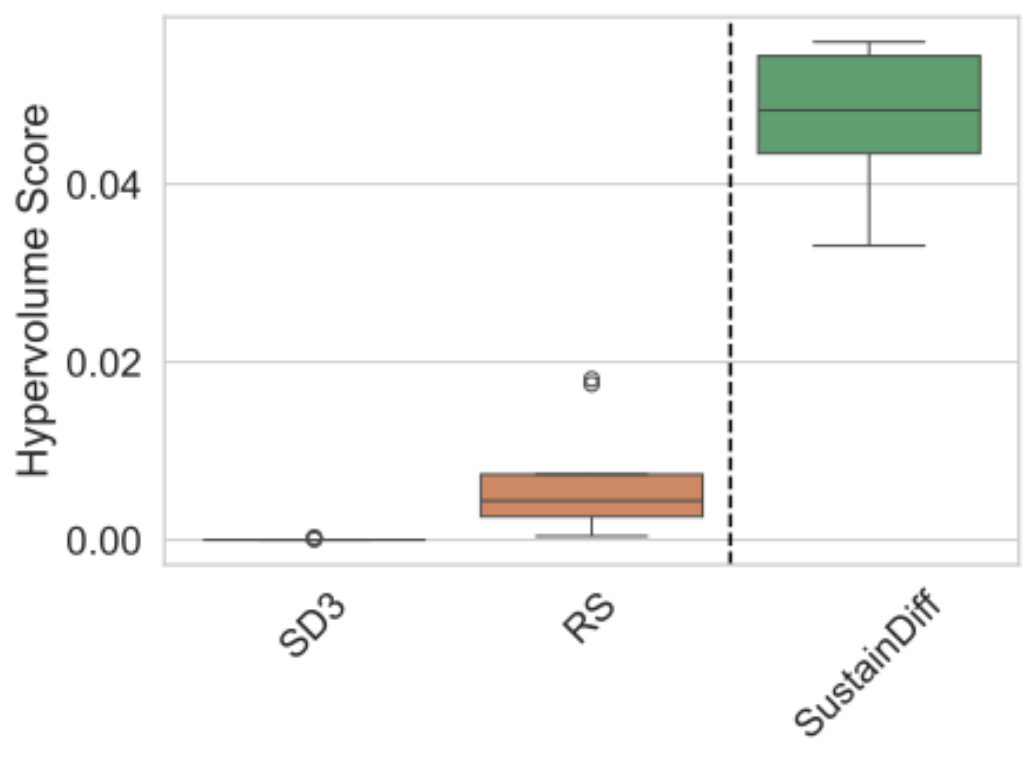
- Single-objective scores
- Trade-offs:
  - Hypervolume: measure how well optimal solutions cover the objective space
  - Pareto optimality: count how many solutions from each strategy are in the Pareto front

Results

RQ1: Baseline Comparison – Single Objectives

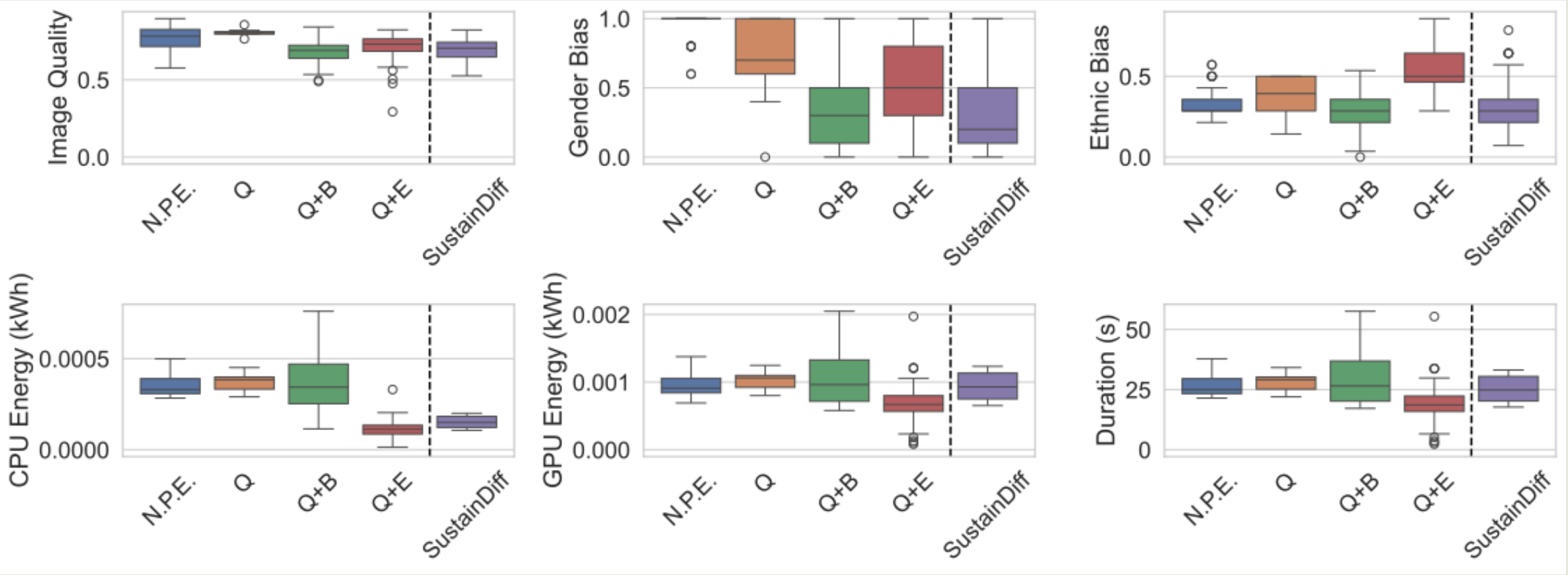


RQ1: Baseline Comparison – Trade-Offs

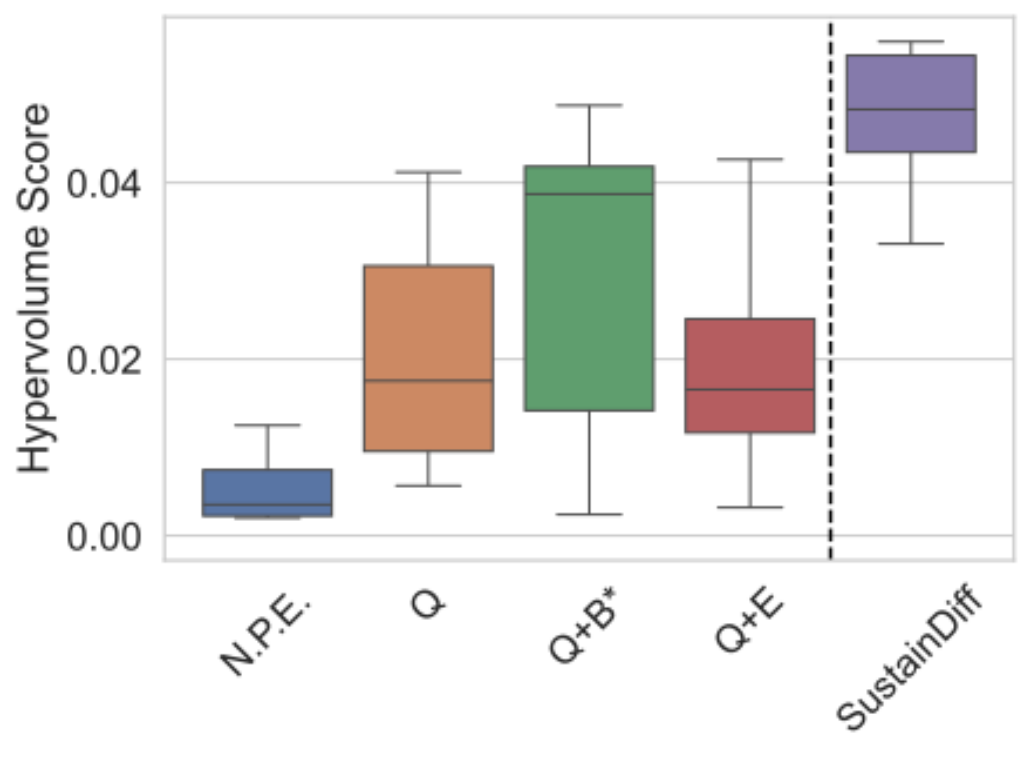


Strategy	Pareto Optimal Solutions
SustainDiffusion	28
Random Search	2
SD3 Default	0

RQ2: Ablation Study – Single Objectives



RQ2: Ablation Study – Trade-off



Strategy	Pareto Optimal Solutions
SustainDiffusion	28
Img. Q. + Bias	21
Img. Q. + Energy	14
No Prompt Eng.	6
Img. Q.	5

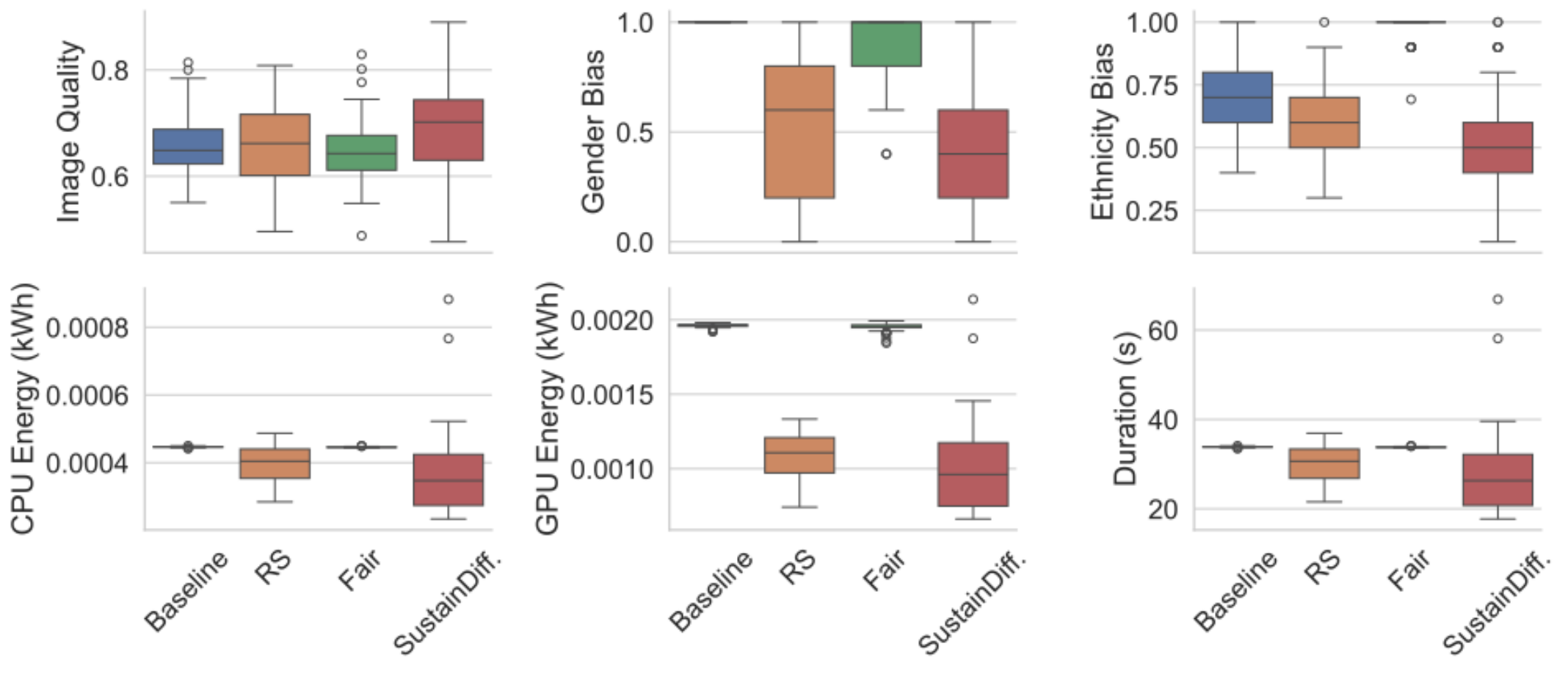
RQ3: Variability - Results

Kruskall-Wallis H test p-value for each objective in ten execution runs

Objective	$p-$value
Image Quality	0.835
Gender Bias	0.057
Ethnic Bias	0.801
CPU Energy	2.545×10^{-17}
GPU Energy	2.749×10^{-17}
Duration	3.383×10^{-17}

The post-hoc Dunn's test reports a statistically significant difference in 36% of the runs for CPU Energy and Duration, and in 38% of the cases for GPU Energy.

RQ4: Generalisation - Results



Limitations

Sustainability of SustainDiffusion

- Challenge: The energy required by a complete run of SustainDiffusion (25 generations) can be recovered with ~4030 prompts of an SD model optimised by SustainDiffusion.

- Solution: Adopt surrogate models as fitness function estimators.

Lack of Qualitative Evaluation

- Challenge: The fitness of the solutions returned by SustainDiffusion is evaluated using quantitative-automated metrics.

- Solution: Perform a user-evaluation on the images generated by an SD model optimised by SustainDiffusion.

Conclusions

Takeaways

- The gender and ethnic bias in SD models is extremely challenging to mitigate, and hyperparameter tuning alone is not enough.
- There is no correlation between gender and ethnic bias and the other objectives optimised by SustainDiffusion.
- CPU energy, GPU energy and Time Duration are highly positively correlated; therefore, optimising for one positively impacts the others.
- With SustainDiffusion, we demonstrate how improving the social and environmental sustainability of SD models is possible without the need for fine-tuning or changing the model architecture.

Future Work

- Apply SustainDiffusion to other image generation models like Flux.1
- Evaluate different search strategies like NSGA3 or Weighted Sum.
- Test the approach on additional prompts for multiple tasks.

Preprint

Thank you for your attention!