

StableZero: Improving General Image Generation Through Visualisation Feedback

Kazim Laos[1], Hector D. Menendez[2(✉)], and Gema Bello-Orgaz[3]

[1] London, UK
[2] King's College London, London, UK
hector.menendez@kcl.ac.uk
[3] Universidad Politecnica de Madrid, Madrid, Spain
gema.borgaz@upm.es

Abstract. Improving image generation systems is essential to reduce compute, energy, and token usage. A key bottleneck is reliance on textual prompts, which are often ambiguous and fail to steer generative models reliably. Prior work, such as StableYOLO, mitigates this via prompt and parameter optimisation with YOLO-derived feedback, but its effectiveness is bounded by YOLO's fixed object taxonomy. This paper introduces StableZero, an enhanced framework that closes this gap by incorporating image captioning into the optimisation loop. Caption-based feedback enables open-vocabulary, semantically grounded search beyond category limits. StableZero couples prompt engineering, semantic fitness scoring, and genetic-algorithm optimisation, building on the StableYOLO architecture. We evaluate on both object-centric and abstract prompts, targeting gains in realism and diversity. Experiments indicate complementary strengths: StableYOLO excels on object-focused prompts, whereas StableZero outperforms on abstract or holistic scenes. Analysis of prompt vocabularies and parameter distributions suggests hybrid strategies can further advance image quality. Overall, captioning-based evaluation offers a more adaptable alternative to category-restricted methods. By aligning feedback with semantics rather than fixed labels, it reduces wasted iterations and improves sample efficiency. StableZero establishes a robust foundation for automated image evaluation and optimisation in creative AI.

Keywords: Generative AI · Text-to-image synthesis · Stable Diffusion · Image captioning · Semantic similarity · Genetic-algorithm optimisation · YOLO · StableYOLO · StableZero

1 Introduction

Generative Artificial Intelligence (AI) refers to computational techniques capable of producing seemingly novel and meaningful content –such as images, text, or

K. Laos—Independent Researcher.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16767, pp. 492–509, 2027.
https://doi.org/10.1007/978-3-032-30491-9_31

audio– based on patterns learned from training data [1]. In the domain of image generation, models such as Stable Diffusion [17], DALL-E, and Midjourney have facilitated the creation of high-quality visuals from textual descriptions, also referred to as prompts [18]. These systems rely on various parameters, including guidance scale, inference steps, and random seed, all of which influence the quality and characteristics of the final output [3].

Improving the quality of generated images remains a central research challenge, particularly in producing photorealistic outputs [2,13]. A major difficulty lies in evaluating image quality, which often depends on subjective human judgment or automated metrics with limited scope, similar to LLMs [5]. Although object-detection models such as YOLO (You Only Look Once) can assist in assessing generated images, their utility is constrained by predefined categories, making it difficult to generalize evaluations for open-ended or highly creative prompts.

A recent approach addressing this issue is StableYOLO, which integrates Stable Diffusion with YOLO and leverages a genetic algorithm to optimize image quality [2]. In this work, we extend this framework by proposing a more flexible mechanism that is not bound by YOLO's restricted classification vocabulary and can instead provide holistic improvements across diverse photorealistic images.

The improved system, named StableZero, introduces an alternative feedback loop based on image captioning. By interpreting visual content in natural language, the system can guide the optimization process in a broader and semantically meaningful way. Its goal is to enhance image quality even when the underlying concepts fall outside YOLO's detection range. To achieve this, StableZero employs captioning –or image-description models– to compare the semantic alignment between the original prompt and the generated caption, using this similarity as the fitness function that drives the optimization.

We validate both StableZero and StableYOLO, finding that they significantly outperform their respective baselines (Wilcoxon $p < 0.01$), with StableYOLO excelling on object-centric prompts and StableZero performing better on more abstract or holistic scenes. Across methods, prompt-word and hyperparameter analyses reveal shared drivers of image quality –such as the importance of *photograph*, *depth*, *color*, and *field*– and similar optimal parameter ranges for inference steps (∼56–57), guidance scale (∼10–12), and guidance rescale (∼0.36–0.47). These observations suggest strong complementarity between the two systems and motivate the exploration of hybrid strategies. Our contributions are as follows:

- An evaluation across 20 prompts with multiple generations, showing that StableZero improves quality by +16.8 points over its baseline (0.459 → 0.627) and that StableYOLO achieves an even larger absolute gain (0.4497 → 0.781).
- A lexical analysis identifying shared positive (*photograph*, *depth*, *field*, *color*) and negative (*illustration*, *art*, *error*, *painting*) prompt terms associated with higher image quality.

- A comparative study demonstrating no significant differences in the distributions of key parameters (inference steps, guidance scale, guidance rescale), despite the models relying on different quality metrics.
- A qualitative characterization of the contexts in which each technique performs best, highlighting complementary strengths and motivating their potential combination for further improvements.

2 Background

StableZero builds on several areas of research, including generative AI for image synthesis, visual analysis and captioning systems, and semantic similarity metrics. This section reviews the current state and limitations of generative image models, with a particular focus on evaluating and improving image quality.

2.1 Generative AI for Image Generation

Generative AI has advanced rapidly, particularly in text-to-image generation. These systems use natural-language prompts to synthesise images, transforming how users interact with creative tools. Models such as DALL-E 2, Stable Diffusion, and DeepMind's Imagen [8] rely on transformer-based architectures trained on large-scale datasets containing millions of image-text pairs [3].

Latent Diffusion Models (LDMs), which operate in compressed latent spaces, have emerged as the dominant architecture for high-quality image generation [16]. Although Generative Adversarial Networks (GANs) once played a central role, diffusion models have largely replaced them due to improved stability and output quality [4]. Another relevant methodology is CLIP-based guidance, which aligns image and text semantics through contrastive learning [11].

StableZero builds on Stable Diffusion [19], an LDM-based model, and introduces a feedback loop that leverages caption analysis to optimise the generation process.

2.2 Visualisation Systems and Captioning Approaches

Traditional visualisation systems such as YOLO are designed for object detection and classification [9]. YOLO has been widely adopted for evaluating generated images because it provides structured numerical feedback. However, its reliance on predefined object categories and photorealistic training data limits its usefulness for diverse or creative generative tasks [2].

To overcome these limitations, this work examines captioning models. Models such as BLIP, CLIP [7], and LLaMA-VID [12] can generate descriptive natural-language captions for images, offering a richer and more contextual understanding of visual content. Unlike YOLO, these systems do not inherently produce numerical scores but a description. StableZero transforms description comparisons into quantitative measures using semantic similarity metrics, constructing a fitness function for optimisation (see Sect. 3).

2.3 Caption Comparison Metrics

Unlike YOLO, which provides direct numerical detection scores based on visible objects, caption-based systems generate free-form text and therefore do not yield an immediate optimisation signal. This requires deriving quantitative evaluations from comparisons between captions. To accomplish this, established metrics from automatic summarisation can be used. Metrics such as BLEU, ROUGE, METEOR, BERTScore, and SummaC have been extensively studied [14], with ROUGE being one of the most widely used. However, ROUGE lacks semantic understanding, making it insufficient for evaluating meaningfully aligned captions.

For this work, SimCSE (Simple Contrastive Learning of Sentence Embeddings) is employed as the primary similarity metric [6]. SimCSE enables robust semantic comparisons between captions and provides a reliable basis for the fitness function in the StableZero optimisation framework.

3 StableZero

This section presents StableZero, a photorealistic visual evaluation system extending StableYOLO [2]. The pipeline integrates prompt engineering, semantic evaluation, and evolutionary optimisation. Its goal is to address classical limitations of object-detection-based feedback –and its particularly poor generalisation to unseen categories and visually complex backgrounds– by coupling a generative image model with a caption-driven semantic fitness function.

The workflow consists of four components: (i) generation of synthetic yet realistic images using Stable Diffusion, (ii) the definition of positive/negative prompt modifier spaces and parameter configurations, (iii) a caption-based fitness function for semantic alignment, and (iv) evolutionary search to optimise prompt and parameter selections. This design enables controlled experimentation while increasing both the variability and quality of generated outputs. Figure 1 illustrates the full pipeline.

Unlike approaches relying on static datasets, StableZero employs a generative pipeline in which image creation is driven by user-defined prompts augmented with structured visual descriptors (positive modifiers), parameters, and constraints (negative modifiers). This enables fine-grained control over properties such as lighting, optics, and visual style.

3.1 Prompt Engineering and Search Space

The search space consists of two types of prompt modifiers –positive and negative– and a set of numerical parameters that influence the quality of images generated by Stable Diffusion. StableZero explores this joint space. It retains the original search space of StableYOLO, but extends the vocabulary of positive and negative modifiers. To define the search space, let Σ^* denote the set of

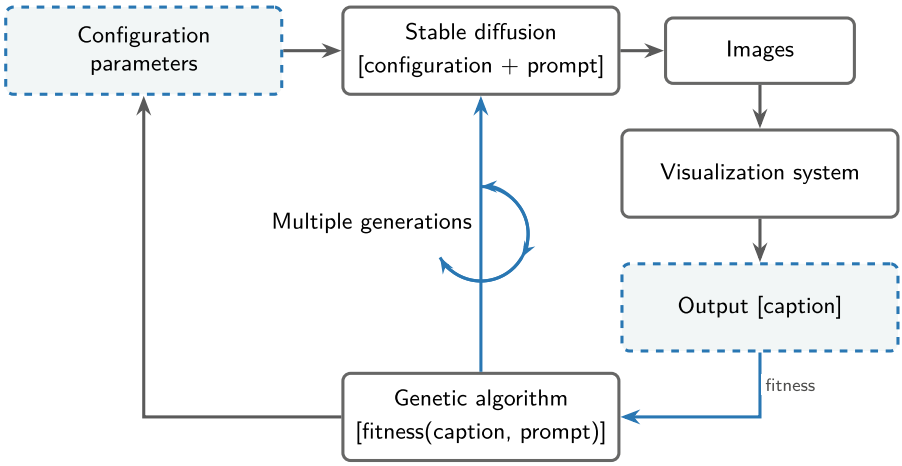

Fig. 1. StableZeropipeline running the four steps to image generation and parameter optimisation.

finite character strings. The user provides a base prompt $p \in \Sigma^*$. We define two ordered modifier sets:

$$\mathcal{P} = (P_1, \ldots, P_m), \qquad \mathcal{N} = (N_1, \ldots, N_n),$$

where $\mathcal{P}$ contains positive modifiers (e.g., lens, film stock, lighting) and $\mathcal{N}$ contains negative modifiers (e.g., "low quality", "blurry").

Binary Subset Selection and Integer Encoding. To keep the search process within a numerical space, a subset of modifiers is selected via the binary expansion of an integer. For a set $\mathcal{S} = (S_1, \ldots, S_\ell)$ and an integer $z \in \{0, \ldots, 2^\ell - 1\}$, let

$$\phi(z, \ell) \in \{0, 1\}^\ell$$

be the ℓ-bit binary vector of z (least/most significant bit convention fixed a priori). The selected subset is

$$\mathrm{Sel}(z; \mathcal{S}) \;=\; \{\, S_i \;\mid\; [\phi(z, \ell)]_i = 1 \,\}.$$

Out-of-range integers are reduced modulo 2^ℓ. This binary number represents a unique selection of words from $\mathcal{P}$ and $\mathcal{N}$.

Prompt Realisation. Given two integers $z^+ \in \{0, \ldots, 2^m - 1\}$ and $z^- \in \{0, \ldots, 2^n - 1\}$, the realised positive and negative prompts are

$$\mathrm{Pos}(p, z^+) \;=\; p \;\|\; \text{“, ”} \;\|\; \Big\|_{S \in \mathrm{Sel}(z^+; \mathcal{P})} S,$$

$$\mathrm{Neg}(z^-) \;=\; \Big\|_{S \in \mathrm{Sel}(z^-; \mathcal{N})} S,$$

where $\|$ denotes string concatenation with commas.

Configuration (phenotype) Space. An individual (phenotype) is the 6-tuple

$$x \;=\; (s,\, g,\, z^{-},\, z^{+},\, r,\, \xi), \quad \text{with}$$

$$s \in \mathbb{N} \quad \text{(inference steps)}, \qquad g \in \mathbb{R} \quad \text{(guidance scale)}, \qquad r \in \mathbb{R} \quad \text{(rescale)},$$

$$z^{-} \in \{0, \ldots, 2^{n}-1\}, \quad z^{+} \in \{0, \ldots, 2^{m}-1\}, \quad \xi \in \mathbb{N} \quad \text{(random seed)}.$$

Box constraints (lower/upper bounds) for (s, g, r) are fixed prior to optimisation, and decoding clips values to these bounds.

3.2 Fitness Function and Semantic Evaluation

Let SD denote Stable Diffusion. For a given phenotype $x = (s, g, z^{-}, z^{+}, r, \xi)$ and a base prompt $p \in \Sigma^{*}$, StableZero produces a batch of M images:

$$\mathcal{I}(x; p) \;=\; \mathrm{SD}\big(\,\mathrm{Pos}(p, z^{+}), \mathrm{Neg}(z^{-});\, s, g, r, \xi, M\,\big) \;=\; \{I_k\}_{k=1}^{M},$$

where the seed ξ determinises the generator.

Let $\mathrm{Cap} : \mathcal{I} \to \Sigma^{*}$ be the captioner [10], and let $f : \Sigma^{*} \to \mathbb{R}^{d}$ be the sentence encoder [6]. In the sentence space, we will employ the cosine similarity,

$$\cos(u, v) \;=\; \frac{\langle u, v \rangle}{\|u\|_2 \, \|v\|_2}.$$

Let $I_k \in \mathcal{I}(x; p)$, the *reference text* is the original base prompt p, and the *generated caption* is $c_k = \mathrm{Cap}(I_k)$. Thus, the per-image semantic score is

$$\varphi_k(x; p) \;=\; \cos\big(f(p),\, f(c_k)\big).$$

The overall fitness is the mean similarity over the batch:

$$F(x; p) \;=\; \frac{1}{M} \sum_{k=1}^{M} \varphi_k(x; p). \tag{1}$$

This replaces YOLO's confidence-based scores with a semantic objective reflecting how well the generated images match the intended prompt.

3.3 Genetic Encoding and Operators

We distinguish the *phenotype* x evaluated by the pipeline and a fixed-length *genotype* $b \in \{0, 1\}^{L}$ manipulated by genetic operators.

Genotype/Phenotype Mapping. The genotype is partitioned as

$$b \;=\; b^{(s)} \,\|\, b^{(g)} \,\|\, b^{(z^-)} \,\|\, b^{(z^+)} \,\|\, b^{(r)} \,\|\, b^{(\xi)},$$

where each block encodes one parameter. Integer fields use binary encoding; real fields use uniform quantisation within fixed bounds:

- $b^{(s)}$ encodes $s \in [s_{\min}, s_{\max}] \cap \mathbb{N}$;
- $b^{(g)}$ encodes $g \in [g_{\min}, g_{\max}]$ with quantisation step Δ_g;
- $b^{(z^-)} \in \{0,1\}^n$ encodes $z^- \in \{0, \ldots, 2^n - 1\}$ via binary;
- $b^{(z^+)} \in \{0,1\}^m$ encodes $z^+ \in \{0, \ldots, 2^m - 1\}$ via binary;
- $b^{(r)}$ encodes $r \in [r_{\min}, r_{\max}]$ with quantisation step Δ_r;
- $b^{(\xi)}$ encodes $\xi \in [0, \xi_{\max}] \cap \mathbb{N}$.

The decoder $D : \{0,1\}^L \to \mathcal{X}$ maps b to x by (i) interpreting integer blocks in binary and (ii) mapping real-valued blocks through uniform quantisation and clamping to the specified bounds. The encoder $E : \mathcal{X} \to \{0,1\}^L$ is defined accordingly.

Two-Point Crossover (Vector-Based). Given two parent genotypes $b^{(1)}, b^{(2)} \in \{0,1\}^L$, choose indices $1 \le c_1 < c_2 \le L$ uniformly at random. The offspring genotype $b^{(\text{off})}$ is formed by exchanging the bit segment between c_1 and c_2:

$$b^{(\text{off})}_{1:c_1-1} = b^{(1)}_{1:c_1-1}, \quad b^{(\text{off})}_{c_1:c_2} = b^{(2)}_{c_1:c_2}, \quad b^{(\text{off})}_{c_2+1:L} = b^{(1)}_{c_2+1:L},$$

with the complementary offspring obtained by swapping the parental roles. Because the genotype concatenates all fields, crossover is *vector-based* at the bit level and can cut within or between fields.

Bit-Flip Mutation. Given a genotype $b \in \{0,1\}^L$, the mutated genotype $\tilde{b}$ is obtained by independent Bernoulli flips at rate $\mu \in (0,1)$:

$$\mathbb{P}\big(\tilde{b}_\ell \neq b_\ell\big) = \mu \quad \text{for each } \ell \in \{1, \ldots, L\}.$$

On the phenotypic prompt selectors, bit flips toggle membership of individual modifiers. On real/integer parameters, bit flips change the quantised value after decoding; decoding D ensures all values respect bounds.

Evolutionary Loop. Let $\mathcal{P}_t$ be the population of genotypes at generation t. The algorithm iterates:

1. Decode $b \in \mathcal{P}_t$ to $x = D(b)$, evaluate $F(x; p)$ via (1).
2. Select parents (e.g., tournament or proportionate selection; choice fixed prior to runs).
3. Apply two-point crossover with probability p_c and bit-flip mutation with rate μ to produce offspring.
4. Form the next population $\mathcal{P}_{t+1}$ using a replacement strategy (e.g., $(\mu+\lambda)$ or generational).

Only the variation operators (two-point crossover and bit-flip mutation) are used; all other evolutionary components (selection, replacement) are standard and fixed by the experimental setup.

4 Experimental Setup

To evaluate the effectiveness of StableZero, the primary objective is to determine whether it can handle cases that are typically challenging for StableYOLO. These cases involve concepts outside YOLO's predefined taxonomy. The evaluation therefore employs a diverse set of prompts combining unfamiliar objects, complex backgrounds, and abstract concepts, enabling a comparison of the complementary strengths of both techniques. The study aims to answer the following research questions:

RQ1: Does StableZero improve image quality compared to the baseline? To address this question, the study applies StableZero to 20 prompts and evaluates whether its generated results are statistically better than those of the baseline. The baseline is formed by the images generated with the prompts with no extra words. Both the quality of StableZero and the baseline are measure with the caption quality metric.

RQ2: How does StableZero's improvement compare to the optimisation achieved by StableYOLO? Here, the same methodology as in RQ1 is applied, but with results compared against StableYOLO to examine whether its improvements follow a similar trend. This includes quantifying both image quality and the degree of optimisation achieved by each system. The baseline quality in this context is measured with the detection quality of YOLO.

RQ3: Which prompt words improve image quality for both StableZero and StableYOLO? This question analyses the top 10 images for each prompt and identifies the most frequent words appearing in successful generations. Aggregating these across all 20 prompts highlights the most influential positive and negative prompt terms.

RQ4: Which parameter selections best improve image quality for both StableZero and StableYOLO? Following the same approach as RQ3, this analysis focuses on Stable Diffusion parameters produced after optimisation, aiming to identify which parameter ranges most consistently yield high-quality results.

4.1 Experimental Protocol, Parameters, and Resources

We evaluated the system using a curated set of 20 complex prompts, more detailed than those used in StableYOLO, and executed all experiments on remote GPU servers to enable large-scale generation and real-time performance monitoring. Although no human data were collected, the use of AI-generated imagery and pre-trained models required institutional ethical assessment, and formal approval was obtained from Solent Southampton University, where the

first author developed the work. All models are publicly available under permissive open licences, and all prompts and captions are synthetic and ensuring GDPR compliance.

The genetic algorithm parameters follow those of the original StableYOLO work to ensure a fair comparison. The number of generations is set to 50. Mutation and crossover probabilities are both 0.2. Elitism is 10, with $\mu = 5$ and $\lambda = 5$. The inner mutation probability is also 0.2. The population size is 25, and tournament selection uses a tournament size of 5.

All experiments were executed on a high-performance computing environment consisting of an AMD EPYC Milan 32-core CPU, 723 GB of RAM, 400 GB SSD storage, and two NVIDIA L40S GPUs with 48 GB VRAM each, running Ubuntu 22.04. This configuration provided sufficient compute capacity for the StableZero pipeline, with GPU resources being particularly crucial for the generative processes.

4.2 List of Prompts Used

Table 1 lists the prompts used for evaluation. The prompts were selected to provide a diverse assessment set covering a wide range of contexts. They include realistic scenarios (e.g., "Older woman baking in a cozy kitchen"), historical figures (e.g., "Abraham Lincoln writing a letter"), and fictional or imaginative scenes (e.g., "Darth Vader in a field of flowers", "Frankenstein's Halloween coffee"). The set also spans varied object categories –humans, animals, and inanimate objects –to assess how YOLO handles typical and unconventional imagery. This diversity ensures that the evaluation does not focus solely on standard cases, but also challenges both systems with uncommon and creative situations.

We extend the prompts with two bag-of-words lists. The first is a positive bag containing 60 candidate augmenters for the positive prompt (e.g., "photograph", "digital", "color", "ultra-real", "film grain"). The second is a negative bag containing 26 candidate terms (e.g., "illustration", "painting", "drawing", "art", "sketch", "low-res"). This design doubles the number of options in the original StableYOLO setup and subsumes all of its original bags-of-words [2].

4.3 List of Models Used

The models used in this study include the Stable Diffusion v2 hosted on Hugging Face, developed by Rombach and Esser [16]. This diffusion-based text-to-image generator is designed primarily for English prompts and released under the CreativeML Open RAIL++-M License. As a Latent Diffusion Model, it operates in a compressed latent space with a fixed OpenCLIP-ViT/H text encoder.

For image captioning, we employ BLIP (Bootstrapping Language–Image Pre-training) [10], which generates natural-language descriptions of images and is well suited for multimodal understanding.

Semantic similarity is computed using SimCSE (Simple Contrastive Learning of Sentence Embeddings) [6], based on BERT-base. SimCSE produces robust

Table 1. List of Prompts for the Evaluation.

ID	Prompt	YOLO classes
0	Darth Vader in a Flower Field	None
1	Elderly Woman Baking in Cozy Kitchen	Person/Dining table
2	Handsome Man in Beachwear with Photo on iPhone 14	Person
3	Electrical Engineer Inspecting Control Panel	Person
4	Abraham Lincoln Writing Letter	Person
5	Woman Standing at a Wedding Cake Table	Cake/Person
6	Cheerful Man in Goofy Hat	Person
7	Happy Man in Flower Field	Person
8	Circus Cat on Telephone Pole	Cat
9	Asian Man Sleeping on Couch	Person
10	Sunset in a Study Prompt	Chair/Table
11	Chinese Woman in Relaxing Yard Scene	Person
12	Aging Jack Russell Terrier Napping	Dog
13	Donkey Kong as a Chimpanzee	None
14	Frankenstein's Halloween Coffee	Cup
15	Putin in Dung	Person
16	Dog Using Laptop	Dog/Laptop
17	Mouse Holding Computer Mouse	Mouse
18	Security Guard and Woman	Person
19	Mafia Couple Ruling Hell	Person

sentence embeddings that enable reliable semantic comparisons between textual descriptions.

Finally, we use Ultralytics' YOLOv8 model for StableYOLO, released on 10 January 2023 [15], which provides state-of-the-art performance across a wide range of object detection tasks.

5 Results

This evaluation compares the performance of StableZero and StableYOLO. The experimental evaluations show that both systems significantly outperform their respective baselines, with each method excelling under different conditions: StableYOLO performs best on object-focused prompts (Sect. 5.2), whereas StableZero achieves higher quality on abstract or holistic prompts (Sect. 5.1). Analyses of prompt vocabulary (Sect. 5.3) and parameter distributions (Sect. 5.4) further reveal shared factors influencing image quality, suggesting that both approaches operate under similar optimal settings. Overall, these findings highlight the complementary nature of the two systems and indicate that combining

their strengths could lead to further improvements in image generation quality (Sect. 5.5).

5.1 RQ1: StableZero Quality

The first experiment evaluates whether StableZero improves image quality compared to its baseline. Using the full set of 20 prompts, each repeated four times over multiple generations, the results (Fig. 2) show that StableZero outperforms the baseline in every case. On average, the improvement is 16.8 points, demonstrating the effectiveness of the approach.

The mean quality score achieved by StableZero is 0.627 ± 0.193, compared to 0.459 ± 0.187 for the baseline. A Wilcoxon test confirms that the improvement is statistically significant, with a p-value of $1.91 \cdot 10^{-6}$, well below the 0.01 threshold used to reject the null hypothesis.

We also examined which prompts were easiest or most challenging for the system. StableZero reached its highest scores on prompts 0 (0.95) and 5 (0.84), both of which involve relatively simple semantic descriptions ("Darth Vader in a Flower Field" and "Woman Standing at a Wedding Cake Table"). These results suggest that BLIP, the captioning model, incorporates cultural knowledge enabling it to describe iconic characters such as Darth Vader. In contrast, prompts 19 (0.22) and 4 (0.24) yielded the lowest scores. These prompts ("Mafia Couple Ruling Hell" and "Abraham Lincoln Writing Letter") focus heavily on specific individuals and rely on contextual or symbolic cues that are harder for the model to interpret.

RQ1 Answer. StableZero improves image quality across all 20 prompts, achieving an average gain of 16.8 points over the baseline.

5.2 RQ2: Comparison Between StableZero and StableYOLO

For StableYOLO, the improvement over its baseline is even greater, partly because its metric focuses directly on image quality and allows more fine-grained optimisation (see Fig. 2). The baseline score is 0.450 ± 0.160, whereas StableYOLO achieves 0.781 ± 0.123. As with StableZero, these improvements are statistically significant according to the Wilcoxon test ($p = 1.91 \cdot 10^{-6}$).

StableYOLO reaches its best performance on prompts 3 (0.90) and 2 (0.89), both dominated by person-related elements (see Table 1). In contrast, its lowest scores appear for prompts 0 (0.42) and 1 (0.61), in which the first provides no viable optimisation path and the second contains multiple competing objects.

A comparison of the images with the highest improvements reveals that StableYOLO and StableZero frequently complement each other. For prompts such as 0, 1, and 5, StableYOLO delivers substantially better results, whereas in other cases StableZero performs more effectively. This pattern arises because StableYOLO excels when there are recognisable objects in the image

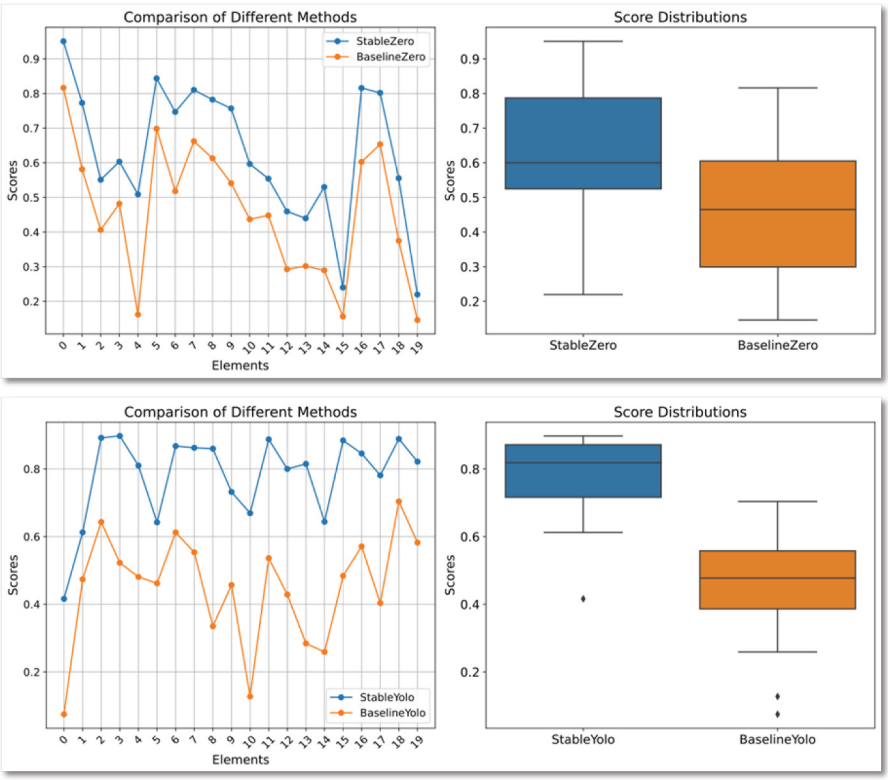

Fig. 2. Comparison between StableZero and its baseline (top), and StableYOLO and its baseline (bottom).

that guide the optimisation process, while StableZero performs better when the scene lacks strong object-level cues. In both situations, the techniques outperform their baselines consistently.

Several prompts –particularly 4 and 19, and to a lesser extent 12 and 13– prove difficult for StableZero, which struggles to achieve meaningful improvement. In these cases, StableYOLO generates noticeably higher-quality outputs. In contrast, those prompts where StableYOLO was able to identify no classes, like Dark Vader's prompt, make it hard to optimize the image quality following StableYOLO's fitness.

Both systems achieve statistically significant improvements over their baselines ($p < 0.01$). This is also visually evident in the result distributions, where a consistent performance gap appears in every run.

RQ2 Answer. StableZero and StableYOLO show complementary strengths. While StableYOLO achieves larger overall improvements, it performs poorly on prompts where StableZero can still enhance quality; conversely, StableYOLO benefits from a clearer optimisation gradient when identifiable objects are present.

5.3 RQ3: Best Words Associated with Image Quality

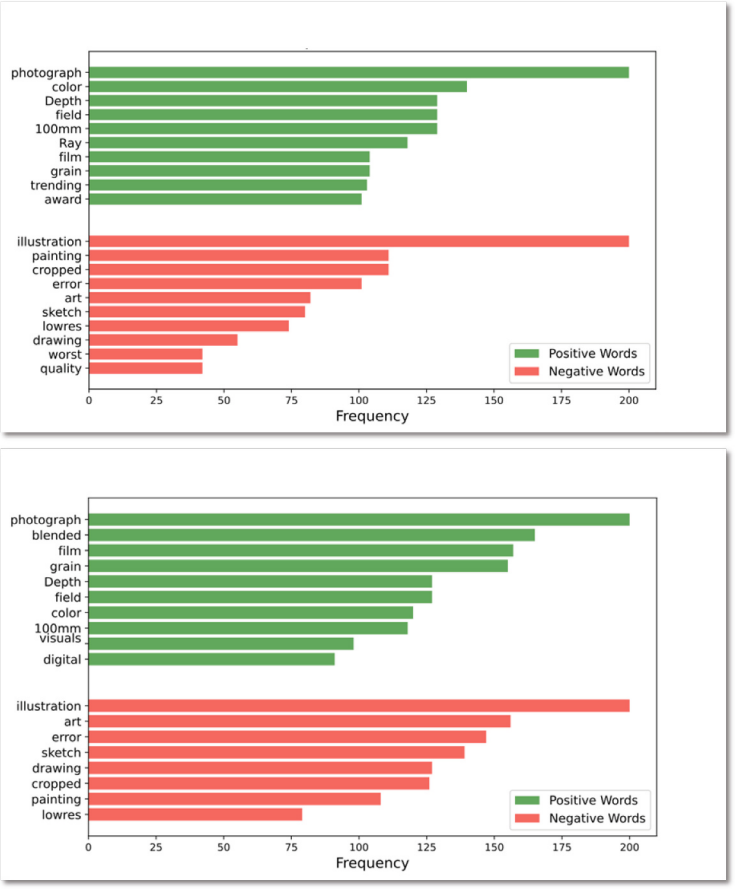

Fig. 3. Most frequent words among the top 10 images for all 20 prompts using StableZero (top) and StableYOLO (bottom).

Many of the most frequent positive-prompt words are shared between the two techniques (see Fig. 3). The word *photograph* appears most often for both. For

StableZero, the next most frequent words are *color*, *depth*, *field*, and *100 mm*, each appearing more than 125 times. For StableYOLO, the leading terms include *blended*, *field*, and *grain* (over 150 occurrences), as well as *depth* and *field* (over 125 occurrences). This pattern indicates that *depth* and *field* are important across both models, while *blended* is specific to StableYOLO.

Other differences include *film*, which is less important for StableZero but remains relevant for StableYOLO. Unique words also appear: StableYOLO favours *visual* and *digital*, whereas StableZero employs *trending* and *award*.

For negative prompts, both models use *illustration* most frequently. StableYOLO exhibits a more even distribution across *art*, *error*, *sketch*, *drawing*, and *crop*. In contrast, StableZero's negative-prompt distribution is more skewed, with a sharp dominance of *illustration*. Some words such as *painting* and *low resolution* appear more often in StableZero, while *quality* and *worst* appear exclusively there, likely reflecting differences in search behaviour rather than aesthetic preferences.

Prompt-length analysis (Table 2) shows that StableZero tends to use longer positive prompts (20.29 words on average) than StableYOLO (17.17 words). For negative prompts the pattern reverses: StableZero uses fewer words (4.49 vs. 5.41). In both cases, Wilcoxon tests confirm statistically significant differences ($p < 0.01$).

Table 2. Average prompt length for StableZero and StableYOLO, with significant differences in both and negative prompts ($p < 0.01$). Parameter comparison between StableZero and StableYOLO (no significant differences according to the Wilcoxon test).

Measure	StableZero	StableYolo	p-value
Average words (positive prompt)	20.29	17.17	7.0×10^{-6}
Average words (negative prompt)	4.49	5.41	2.1×10^{-6}
Inference steps (*mean* $\pm$ SD)	56.77 ± 23.83	56.44 ± 27.24	0.87
Guidance scale (*mean* $\pm$ SD)	10.07 ± 4.80	11.68 ± 5.81	0.43
Guidance rescale (*mean* $\pm$ SD)	0.47 ± 0.27	0.36 ± 0.24	0.11

RQ3 Answer. Overall, both systems prioritise similar positive prompt words—*photograph*, *depth*, *color*, and *field* –and similar negative words– *illustration*, *art*, *error*, and *painting*.

5.4 RQ4: Parameter Selection and Image Quality

RQ4 examines how different Stable Diffusion parameters affect image quality. A notable finding is that both StableZero and StableYOLO converge to very

similar parameter distributions, despite using different evaluation criteria. The three parameters studied are the number of inference steps, the guidance scale, and the guidance rescale (see Table 2).

For inference steps, StableZero uses 57.77 ± 23.83 on average, while StableYOLO uses 56.44 ± 27.24—nearly identical values, though StableYOLO shows more variability. A Wilcoxon test indicates no significant difference ($p = 0.87$).

The guidance scale also proves similar across models: 10.07 ± 4.80 for StableZero and 11.68 ± 5.81 for StableYOLO ($p = 0.43$). Likewise, guidance rescale exhibits no significant difference: 0.47 ± 0.27 vs. 0.36 ± 0.24 ($p = 0.11$).

These results show that both techniques operate under nearly identical optimal parameter settings. Despite different optimisation heuristics and fitness functions, both approaches gravitate toward similar stable regions of the parameter space, suggesting potential for cross-method improvement.

RQ4 Answer. Both systems converge to near-identical parameter ranges: ∼[56,57] inference steps, ∼[10,11] guidance scale, and ∼[0.36,0.47] guidance rescale.

5.5 Discussions

The highest-performing StableYOLO prompts (Fig. 4), such as "Electrical Engineer Inspecting Control Panel" and "Chinese Woman in Relaxing Yard Scene," succeed because they provide strong object-level cues that guide the optimisation. In contrast, StableZero produces more holistic images by focusing on the entire scene.

Weaker StableYOLO prompts –such as "Darth Vader in a Flower Field" and "Sunset in a Study Prompt"– perform poorly when multiple objects interact or when object detection fails. In these scenarios, StableZero produces more coherent and detailed images.

Similarly, StableZero works best on prompts like "Elderly Woman Baking in Cozy Kitchen" and "Woman Standing at a Wedding Cake Table," where global scene coherence is crucial. Meanwhile, StableYOLO tends to emphasise recognisable objects.

Weaker StableZero prompts –such as "Abraham Lincoln Writing Letter" and "Mafia Couple Ruling Hell" – are difficult due to their abstract or symbolic nature, resulting in unrealistic or blurry outputs. Here, StableYOLO performs somewhat better by grounding its optimisation in detectable objects (Fig. 5).



Fig. 4. Comparison of StableYOLO's best and worst prompts (above) with corresponding StableZero outputs (below).



Fig. 5. Best and worst prompts from StableZero (below) with corresponding StableYOLO outputs (above).

6 Conclusions and Future Work

This work presents StableZero, a substantial extension of StableYOLO that replaces YOLO's restricted category-based feedback with a caption-driven evaluation mechanism. StableZero proved effective in producing more flexible assessments and improving image quality in tasks where YOLO encountered limitations. The evaluation shows that StableZero and StableYOLO offer complementary strengths: StableYOLO performs best on object-focused prompts, while StableZero excels on abstract or open-ended scenes. This suggests that combining both approaches could yield a more robust solution than either system alone.

Several avenues for future work emerge from this evaluation. More advanced captioning models could be explored to reduce noise in the feedback loop, and more efficient optimisation strategies may help mitigate the computational cost associated with genetic algorithms. Hybrid systems that integrate both captioning and object detection could also provide a stronger balance between precision and flexibility.

Additional opportunities for extending StableZero include testing it on larger and more diverse datasets, incorporating caption-based evaluation directly into model training rather than solely during inference, and investigating human-in-the-loop methods that combine automated scoring with user preference signals. These extensions would not only strengthen the technical performance of the system but also better align it with human perceptions of creativity and quality.

In conclusion, StableZero demonstrates that caption-driven evaluation is a promising direction for generative image AI. While further refinement is still needed, this work lays a solid foundation for developing more adaptable, scalable, and human-aligned evaluation methods in the future.

Acknowledgments. The support of the UKRI Trustworthy Autonomous Systems Hub (reference EP/V00784X/1), Trustworthy Autonomous Systems Node in Verifiability (reference EP/V026801/2) and the Comunidad Autónoma de Madrid under ALENTAR-J-CM project (reference TEC-2024/COM-224) is gratefully acknowledged.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Archana Balkrishna, Y.: An analysis on the use of image design with generative AI technologies. Int. Trend Sci. Res. Devopment **8**(1), 596–599 (2024). https://www.ijtsrd.com/papers/ijtsrd63468.pdf
2. Berger, H., et al.: Stableyolo: optimizing image generation for large language models. In: Arcaini, P., Yue, T., Fredericks, E.M. (eds.) SSBSE 2023. LNCS, vol. 14415, pp. 133–139. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-48796-5_10
3. Bolya, D., Hoffman, J.: Token merging for fast stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4599–4603 (2023). https://doi.org/10.1109/CVPRW59228.2023.00484
4. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: an overview. IEEE Signal Process. Mag. **35**(1), 53–65 (2018). https://doi.org/10.1016/j.jjimei.2020.100004
5. Del Barrio, M., Laos, K., Lara, M.J.V., García, C.G., Menendez, H.D.: Size doesn't matter: Assessing the trustworthiness of large language models in medical contexts: a focus on epidural information retrieval. Artif. Intell. Med. 103379 (2026). https://doi.org/10.1016/j.artmed.2026.103379

6. Gao, T., Yao, X., Chen, D.: SimCSE: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6894–6910 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.552
7. Hafner, M., et al.: Clip and complementary methods. Nature Rev. Meth. Primers **1**(1), 20 (2021). https://doi.org/10.1101/2020.10.06.327445
8. Jamal, S., Wimmer, H., Rebman Jr, C.M.: Perception and evaluation of text-to-image generative AI models: a comparative study of dall-e, google imagen, grok, and stable diffusion. Issues Inf. Syst. **25**(2), 277–292 (2024). https://doi.org/10.48009/2_iis_2024_123
9. Jiang, P., Ergu, D., Liu, F., Cai, Y., Ma, B.: A review of yolo algorithm developments. Procedia Comput. sci. **199**, 1066–1073 (2022). https://doi.org/10.1016/j.procs.2022.01.135
10. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900. PMLR (2022). https://doi.org/10.5555/3618408.3619222
11. Li, Y., et al.: Supervision exists everywhere: a data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 (2021). https://doi.org/10.48550/arXiv.2110.05208
12. Li, Y., Wang, C., Jia, J.: Llama-vid: an image is worth 2 tokens in large language models. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024. LNCS, vol. 15104, pp. 323–340. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72952-2_19
13. Menendez, H.D., Bello-Orgaz, G., Atencia, C.R.: Deepstableyolo: Deepseek-driven prompt engineering and search-based optimization for ai image generation. In: XVI Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (2025). https://openreview.net/forum?id=hZucDPawRu
14. Menendez, H.D., Dakhama, A.: Automatic summarization evaluation: methods and practices. In: Menéndez, H.D., et al. (eds.) ICTSS 2024. LNCS, vol. 15383, pp. 169–181. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-80889-0_12
15. Rahim, A., Yuan, F., Barabady, J.: An ultralytics yolov8-based approach for road detection in snowy environments in the arctic region of norway. Comput. Mater. Continua **83**(3) (2025). https://doi.org/10.32604/cmc.2025.061575
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022). https://doi.org/10.1109/CVPR52688.2022.01042
17. Tang, R., et al.: What the DAAM: interpreting stable diffusion using cross attention. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5644–5659 (2023). https://doi.org/10.18653/v1/2023.acl-long.310
18. Tansri, F.F., Monem, N., Weinberg, L.: Authenticity in biased diversity: Investigating the language of prompt performances in ai image generators. J. Aesthet. Creat. Art Manag. **4**(1), 75–101 (2025). https://doi.org/10.59997/jacam.v4i1.5414
19. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S., Kim, J.: Text-to-image diffusion models in generative AI: a survey. arXiv preprint arXiv:2303.07909 (2023). https://doi.org/10.48550/arXiv.2303.07909