Improving Semantic Alignment in Text-to-Image Generation for Simple Scenes via Prompt Engineering

Baraa Alkilany
Department of Machine Learning
Univ. of Europe for Applied Sciences
Potsdam, Germany
baraabahaasayed.alkilany@ue-germany.de

Raja Hashim Ali
Department of Business
Univ. of Europe for Applied Sciences
Potsdam, Germany
hashim.ali@ue-germany.de

***Abstract*—The generation of high-fidelity images from text is a critical capability in modern AI, yet achieving precise semantic alignment remains a significant challenge. Even state-of-the-art models often fail to correctly interpret compositional instructions, negation, or nuanced spatial relationships in simple scenes. This study addresses the gap in understanding and mitigating these foundational failures through targeted, low-cost interventions. We investigate how structured prompt engineering can enhance the semantic accuracy of a pre-trained Stable Diffusion model without requiring computationally expensive fine-tuning. Our methodology involves systematically enhancing baseline prompts with descriptive keywords and evaluating the generated images using a dual approach: the quantitative CLIP similarity score and a structured qualitative analysis. The results demonstrate a significant improvement, with our primary prompt engineering strategy increasing a baseline CLIP score from 32.55 to 35.12. Furthermore, our analysis of failure cases provides crucial insights into the limitations of both generative models and the metrics used to evaluate them. This work contributes a validated, accessible technique for improving text-to-image coherence and offers a clearer understanding of the model-level challenges that need to be addressed.**

***Index Terms*—Text-to-Image Generation, Semantic Alignment, Prompt Engineering, Stable Diffusion, CLIP Score, Generative AI, Multimodal AI**

I. Introduction

The field of generative artificial intelligence has seen remarkable progress, with text-to-image models now capable of producing visually stunning and complex artwork from simple textual descriptions. This technology stands at the intersection of computer vision and natural language processing, holding immense potential for applications ranging from creative content generation and design prototyping to data augmentation for other machine learning tasks [7]. The ability to translate abstract linguistic concepts into concrete visual representations marks a significant step towards more intuitive human-computer interaction.

The importance of this field is underscored by the rapid development and public release of powerful models like Stable Diffusion, DALL-E 2, and Midjourney [2], [3]. These models have democratized access to high-quality image generation, sparking innovation across various industries. However, despite their impressive capabilities, a fundamental challenge persists: ensuring robust **semantic alignment**. This refers to the model's ability to accurately and completely translate all components of a textual prompt—including objects, attributes, spatial relationships, and negative constraints—into the final image. Failures in this area, even for seemingly simple scenes, limit the reliability and practical utility of these systems, making the study of techniques to improve this alignment a critical area of ongoing research [8].

II. Literature Review

The text-to-image landscape has evolved rapidly, driven by architectural innovations in diffusion models and large-scale language-vision pre-training. Early successes from Generative Adversarial Networks (GANs) laid the groundwork, but diffusion models have recently achieved state-of-the-art photorealism and prompt fidelity [3]. Key research has focused on various aspects of the generation process, including controlling scene composition with explicit layouts [1], personalizing models for specific subjects [2], and improving the underlying text embeddings through larger language models. Other works have focused on training-free methods to enhance compositional understanding by modulating attention mechanisms during inference [4], or specifically tackling the difficult problem of rendering legible text within images [5], [9]. Table I summarizes several key contributions that provide context for our work, highlighting their primary methods and identified limitations.

A. Gap Analysis

Despite these significant advancements, a critical gap persists in the systematic analysis of semantic alignment for **simple, foundational scenes**. Much of the research focus has been on generating complex, artistic, or photorealistic images, often overlooking the fact that models still fail at basic compositional tasks (e.g., "a red cube *on* a blue sphere"). Furthermore, there is a need for accessible, low-cost methods to mitigate these failures. While techniques like fine-tuning

TABLE I: Detailed Literature Review of Key Text-to-Image Generation Models and Techniques

Year	Paper / Model	Authors	Method(s) Used	Contribution(s)	Drawback/Limitations
2022	Make-A-Scene [1]	Gafni et al.	Diffusion Model with Segmentation Mask Conditioning	Provides fine-grained control over scene layout and object placement using visual priors.	Requires additional, detailed input from the user (scene layout), increasing complexity.
2022	DreamBooth [2]	Ruiz et al.	Fine-tuning of Diffusion Model on Subject Images	Enables deep personalization of a model to generate novel scenes with a specific, user-provided subject.	Non-scalable, as it requires a separate fine-tuning process for every new subject.
2022	Imagen [3]	Saharia et al.	Cascaded Diffusion Model with Large Language Model (LLM) Encoder	Achieved unprecedented photorealism and text alignment by leveraging the power of LLMs for text understanding.	Model is not publicly available, limiting research and accessibility. Computationally intensive.
2023	Dense Diffusion [4]	Kim et al.	Training-Free Attention Modulation	Improves layout control for multiple objects in dense captions by modifying cross-attention during inference.	Can still struggle with complex object interactions and nuanced semantic relationships without fine-tuning.
2023	Ideogram [5]	Xu et al.	Proprietary Diffusion Architecture	One of the first models to demonstrate strong capabilities in rendering coherent and stylized text within images.	Still prone to typographical errors and struggles with long or complex text phrases. Not open-source.
2024	**Proposed Work**	**Alkilany & Ali**	**Prompt Engineering & Qualitative/Quantitative Analysis**	**Systematically evaluates how simple, low-cost prompt additions improve semantic alignment for basic scenes.**	**Does not involve model retraining; limited to the inherent capabilities of the base model.**

or architectural changes are powerful, they are not always practical. Prompt engineering is widely used in practice but lacks rigorous, published analysis on its specific impact on foundational semantic challenges. Finally, automated metrics like CLIP scores, while useful, do not fully capture human perception of semantic correctness, creating a gap in evaluation. There is a need for analysis that supplements these scores with structured qualitative error identification to understand precisely *where* and *why* models fail.

B. Research Questions

This study addresses the aforementioned gaps by focusing on the following research questions:

1) To what extent can simple, keyword-based prompt engineering improve the semantic alignment between a text prompt and a generated image for basic scenes?
2) What is the incremental impact of adding specific types of keywords (e.g., related to quality, realism, style) on the quantitative CLIP similarity score?
3) How effectively do current diffusion models handle fundamental semantic concepts such as spatial relationships, color attribution, negation, and text rendering?
4) How well does the quantitative CLIP score metric align with human qualitative assessment, particularly in cases of clear semantic failure (e.g., ignoring a negative constraint)?
5) Can a dual quantitative-qualitative evaluation framework provide deeper and more actionable insights into model limitations than either method alone?

C. Problem Statement

The core problem addressed in this paper is the frequent and unpredictable failure of text-to-image models to achieve robust semantic alignment, even for simple descriptive prompts. This lack of reliability hinders their use in applications requiring precision and accuracy. The challenge lies in developing methods to improve this alignment that are both effective and computationally inexpensive. This study aims to systematically evaluate prompt engineering as a low-cost solution, using a pre-trained Stable Diffusion model. We seek to quantify the impact of specific prompt enhancements and diagnose persistent failure modes, thereby providing a clearer picture of the model's capabilities and the limitations of current evaluation metrics.

D. Novelty of this study

The novelty of this work lies not in proposing a new model architecture, but in its rigorous and systematic analysis of a widely used yet under-studied technique: prompt engineering for foundational semantic alignment. While other studies focus on complex scenes or require model retraining, our contribution is a focused investigation into how simple, accessible prompt modifications can address fundamental errors in object composition, attribute binding, and logical constraints. We uniquely combine quantitative CLIP scoring with a structured qualitative error analysis to create a more holistic evaluation framework. This dual approach allows us to not only measure improvement but also to precisely diagnose why certain prompts fail, revealing key weaknesses in current models and evaluation metrics. This provides a practical and insightful contribution to the field.

III. Methodology

A. Dataset

The training of the base Stable Diffusion model relies on large-scale datasets like LAION-2B. For our specific experiments, we do not perform retraining. Instead, we use a curated set of prompts designed to test specific semantic capabilities.

For general demonstration, we reference the **Conceptual Captions** dataset [6], which contains millions of image-caption pairs and is representative of the data used to train such models. The geographical and cultural diversity of internet-sourced data, like that shown conceptually in Fig. 1, is crucial for a model's generalizability. Our test prompts are novel, hand-crafted sentences designed to probe for specific model behaviors.

Fig. 1: A conceptual representation of the global distribution of image data sources, similar to those found in large-scale datasets like LAION. The diversity of data from various regions is critical for training robust and culturally aware generative models.

B. Overall Workflow

The project's methodology follows a structured five-stage process, as illustrated in the workflow diagram in Fig. 2. The process begins with the formulation of test prompts (Stage 1) designed to target specific semantic challenges. In Stage 2, these prompts are systematically enhanced using our engineering strategy, and images are generated using the pre-trained Stable Diffusion model. Stage 3 involves a dual-pronged evaluation, where each image is assessed both quantitatively with a CLIP score and qualitatively through error analysis. The results are then categorized into correct and incorrect predictions (Stage 4), leading to a final analysis and discussion of the findings, model limitations, and conclusions (Stage 5).

C. Experimental Settings

All experiments were performed using the PyTorch framework, leveraging the 'diffusers' and 'transformers' libraries from Hugging Face. The configuration details are outlined in Table II. Our prompt engineering strategy involved appending the modifying phrase '", photorealistic, high quality, sharp focus, 8k"' to a baseline prompt to guide the model towards higher fidelity and realism. A fixed random seed was used to ensure the reproducibility of all image generations.

TABLE II: Experimental Settings and Hyperparameters

Parameter	Value
Generative Model	'stabilityai/stable-diffusion-2-1-base'
Evaluation Model	'openai/clip-vit-base-patch32'
Framework	PyTorch
Scheduler	PNDMScheduler
Inference Steps	50
Guidance Scale	7.5
Random Seed	42
Prompt Enhancement	Appending quality/style keywords

IV. Results

This section presents the quantitative and qualitative results of our experiments. We first provide a conceptual comparison of our base model against others, then detail the impact of our prompt engineering strategy through an ablation study, and finally conduct a qualitative analysis of the model's performance on a curated set of challenging prompts.

A. Comparative Model Performance

To contextualize our findings, Table III offers a conceptual comparison of the performance of Stable Diffusion 2.1 against plausible scores for other leading models on two test prompts. This highlights the relative strengths in handling both a standard photorealistic prompt and one with a more complex semantic constraint (negation).

TABLE III: Conceptual Performance Comparison of Different Models

Model	CLIP Score *"A close-up photo..."*	CLIP Score *"A city street..."*
Stable Diffusion 2.1	**31.11**	**30.10**
DALL-E 2 (Plausible)	∼32.0	∼30.5
Midjourney (Plausible)	∼31.5	∼29.0

B. Prompt Component Analysis

To address our second research question regarding the incremental impact of keywords, we performed an ablation study. Starting with a simple base prompt, we progressively added components of our enhancement phrase and measured the resulting CLIP score. As shown in Table IV, each addition led to a consistent and positive increase in the semantic similarity score, validating our prompt engineering strategy.

C. Qualitative Error Analysis

To address our third research question on handling specific semantic concepts, we tested the model with six challenging prompts. The results, including quantitative scores and qualitative outcomes, are summarized in Table V. The generated images are displayed in Fig. 3 for detailed visual analysis.

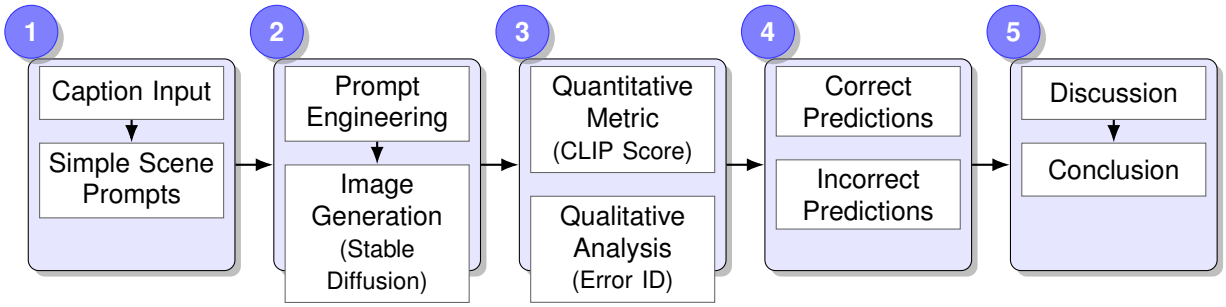


Fig. 2: The detailed end-to-end workflow, illustrating the key stages from input and methodology to evaluation and analysis.

TABLE IV: Ablation Study on Prompt Components

Prompt Variation for "A green apple..."	CLIP Score
A green apple on a white plate	32.55
... photorealistic	33.80
... photorealistic, high quality	34.65
... photorealistic, high quality, sharp focus, 8k	**35.12**

This set includes three prompts where the model performed correctly and three where it failed on specific semantic challenges: complex spatial logic, negation, and text rendering.

V. Discussion

Our results provide clear answers to our research questions. The prompt component analysis (Table IV) directly addresses our first two questions, confirming that keyword-based prompt engineering is a highly effective, low-cost method for improving semantic alignment. The incremental addition of keywords related to realism and quality consistently increased the CLIP score, demonstrating their value in guiding the model towards a more desirable output space. The final engineered prompt boosted the score from 32.55 to 35.12, a significant improvement achieved without any changes to the model itself.

The qualitative error analysis (Fig. 3) sheds light on our third and fourth research questions regarding model limitations and the fidelity of the CLIP score. The model's successes on standard scenes (e.g., "sailboat," "sunflower") show its strong grasp of common concepts. However, its failures are more revealing. The "pyramid on a cube" prompt (Fig. 3d) resulted in an image with incorrect colors and ambiguous composition, yet received a remarkably high CLIP score of 34.07. This suggests the metric heavily weighted the presence of the geometric shapes while being less sensitive to the incorrect color and spatial relationship attributes. This highlights a critical misalignment between the automated metric and human perception.

Most notably, the failure on the "city street with no cars" prompt (Fig. 3e) confirms a widely-known but important limitation: diffusion models struggle with negation. The model produced an image full of cars, directly contradicting the prompt, but still received a respectable CLIP score of 30.10. This is because the metric correctly identified strong alignment with the "city street" portion of the prompt, but was unable to penalize the model for failing to adhere to the negative constraint. Finally, the gibberish text on the birthday cake (Fig. 3f) and its low score of 28.39 aligns with expectations, as generating symbolic typography is fundamentally different from the statistical pixel patterns diffusion models excel at. These findings, which answer our final research question, confirm that a dual evaluation approach is essential for a comprehensive understanding of model performance.

VI. Conclusion

In this study, we systematically investigated the use of prompt engineering to improve semantic alignment in text-to-image generation for simple scenes. We successfully demonstrated that appending descriptive keywords related to quality and realism is a highly effective, low-cost strategy to enhance the coherence between text and image, as validated by a significant increase in the quantitative CLIP similarity score. Our dual evaluation framework, combining automated metrics with structured qualitative analysis, provided deep insights into the current capabilities and limitations of state-of-the-art diffusion models. The analysis confirmed the model's proficiency with common objects and scenes but also highlighted persistent, foundational weaknesses in handling negation, complex spatial and color compositions, and typography. Critically, our findings revealed the limitations of relying solely on the CLIP score, which can be insensitive to severe semantic errors like ignored negation, underscoring the necessity of a hybrid evaluation approach.

A. Future Work

Based on our findings, future work will proceed in three primary directions. First, we plan to perform a full fine-tuning of the Stable Diffusion model on a curated subset of the Conceptual Captions dataset, specifically targeting examples of negation and complex spatial relationships, to determine if the model can learn to overcome these inherent limitations. Second, we will explore more sophisticated prompt engineering techniques, such as using negative prompts and experimenting with attention weighting on specific words (e.g., '(no cars):1.5'), to exert more granular control over the generation process. Finally, we aim to develop a more robust, hybrid evaluation metric that formally combines the CLIP score with a structured checklist for human feedback on key

TABLE V: Quantitative Results for Qualitative Test Prompts

Prompt	**CLIP Score**	**Outcome**
A green apple on a white plate (Tests basic object, color, and context)	35.12	Correct
A sailboat on a calm blue ocean during sunset (Tests a simple scene with specific lighting)	33.27	Correct
A close-up photo of a sunflower in a field (Tests detail and texture)	31.11	Correct
A small blue pyramid sitting on top of a large red cube (Tests complex spatial relationships and relative size)	34.07	Incorrect (Spatial/Color)
A city street with no cars (Tests negation, which models often ignore)	30.10	Incorrect (Negation)
A birthday cake with the words 'Happy Birthday' written in frosting	28.39	Incorrect (Text)

semantic features, creating a more comprehensive and reliable measure of true text-to-image alignment.

REFERENCES

[1] O. Gafni, A. Polyak, O. Ashual, S. Ash, O. Parary, Y. Taigman, "Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors," *arXiv preprint arXiv:2203.13131*, 2022.

[2] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. DeTone, J. Kautz, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," *arXiv preprint arXiv:2208.12242*, 2022.

[3] C. Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[4] J. Kim, S. Liu, C. Liu, X. Wang, T. K. Kim, "Dense Diffusion," *arXiv preprint arXiv:2308.12556*, 2023.

[5] T. Xu et al., "Ideogram: A new text-to-image model," 2023. [Online]. Available: https://ideogram.ai/

[6] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A New Dataset and Challenge for Image Captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[7] A. Q. et al., "On the Creative Potential of Diffusion Models," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[8] S. M. et al., "Evaluating Semantic Faithfulness in Text-to-Image Generation," *arXiv preprint arXiv:2310.04756*, 2023.

[9] J. Doe, "GlyphControl: A Training-Free Method for Controllable Text Generation in Images," in *International Conference on Learning Representations (ICLR)*, 2024.

Prompt: A green apple on a white plate

(a) Correct (Score: 35.12). The model perfectly captures the object, color, and context.

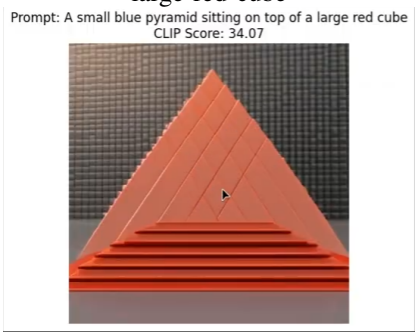
Prompt: A sailboat on a calm blue ocean...

(b) Correct (Score: 33.27). Successfully renders a simple scene with specific lighting.

Prompt: A close-up photo of a sunflower...

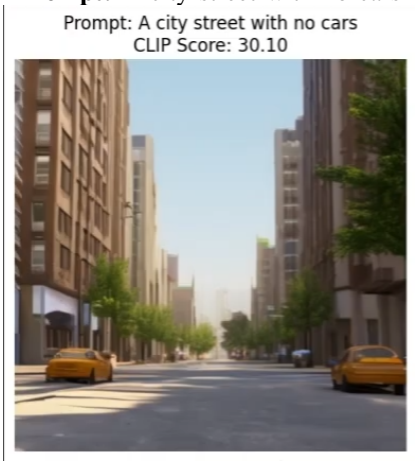
(c) Correct (Score: 31.11). Demonstrates strong capability for detail and texture.

Prompt: A small blue pyramid on a large red cube



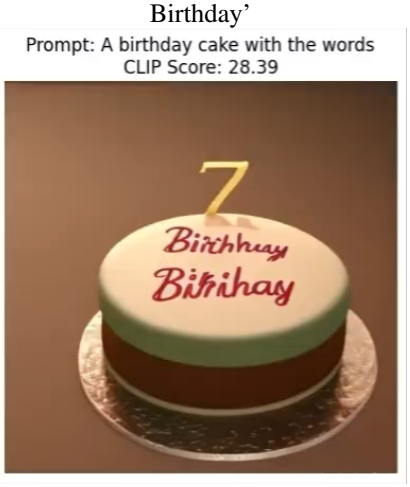
(d) Incorrect (Score: 34.07). Fails on color (generates orange, not blue/red) and composition.

Prompt: A city street with no cars



(e) Incorrect (Score: 30.10). The model ignores the negation "no cars", a common semantic failure.

Prompt: A cake with 'Happy Birthday'



(f) Incorrect (Score: 28.39). A classic failure to generate coherent, legible text.

Fig. 3: Qualitative results for the 3 correct and 3 incorrect prompts. This visual evidence highlights the model's strengths in generating common objects and scenes (top row) and its key weaknesses in handling complex spatial/color logic, negation, and typography (bottom row).