Evaluating Semantic Preservation and Style Drift in Text-to-Image Generation: A BLIP and Stable Diffusion Pipeline with Caption Modification Analysis

Jiafeng Wu, Meiyan Liu

Abstract

Text-to-image models like Stable Diffusion are becoming popular for creative tasks. In this project, we built a pipeline to test how different types of caption modifications—like adding stylistic words or random noise—affect what the model generates and how those outputs are later interpreted. We wanted to examine how well the model handle changes in caption wording, such that when it comes to preserving meaning or interpreting style. We started with real images from the CIFAR-10 and CelebA datasets, used BLIP to generate captions, and then passed those into Stable Diffusion to create new images. We also created modified versions of the captions by adding extra adjectives or random noises, and generated another set of images from those. To see how meaning changes, we used BLIP again to caption all the generated images. We compared the original and modified versions using Rouge, CLIP and Inception scores to measure how much the images and meanings changed. Our results show that Stable Diffusion is fairly sensitive to wording changes, which suggests further implications for how these models understand and translate language into images.

1 Introduction

Recent progress in multimodal machine learning has driven major advances in text-to-image generation including the process of synthesizing images from natural language prompts. Models like Stable Diffusion (SD) have demonstrated great abilities in producing high quality, stylized visuals across a wide range of input texts. However, as these systems are increasingly used in real world applications, questions about how reliably they preserve the intended meaning of the input text, particularly when prompts are modified remain.

This project explores how variations in caption phrasing affect the outputs of a text-to-image generation process. We use Bootstrapped Language-Image Pretraining (BLIP) to generate initial captions from real images in two contrasting datasets: CIFAR-10, which consists of object-centric scenes, and CelebA, which focuses on human faces. These captions are further used as prompts for Stable Diffusion, which generates corresponding synthetic images. We then recaption the generated outputs using BLIP to assess semantic consistency. To test the model's sensitivity to language, we apply two types of caption alterations: stylistic additions (e.g., "professional photo," "animated version") and semantic noise (e.g., "random objects," "chaotic texture"). By comparing the original and altered generations using Rouge, CLIP and Inception scores, we evaluate how changes in texts affect the generated content.

Understanding the relationship between caption phrasing and image generation is important for several reasons. In real world applications such as content creation, design, and human-centered technologies, semantic accuracy is often important. Even small changes in prompt language can lead to significantly different outputs, especially when the prompts are ambiguous or stylistically complex. As text-to-image models are used in more sensitive or creative way, it becomes increasingly important

NeurIPS 2023 Generative AI and Biology Workshop.

to evaluate their robustness to linguistic variation and to understand how language influences visual output. This project contributes to it by providing a structured framework for analyzing prompt sensitivity and semantic alignment in multimodal systems.

2 Methodology

To test how variations in caption phrasing affect text-to-image generation, we conducted a procedure combining image captioning, text-to-image synthesis, and semantic evaluation. Our approach involves three steps - caption generation, image synthesis, and caption re-evaluation.

2.1 Dataset Preparation

For our experiments, we used two widely used datasets: CIFAR-10 and CelebA. The CIFAR-10 dataset comprises 60,000 color images across 10 classes of everyday objects (e.g., airplane, dog, truck), while the CelebA dataset contains over 200,000 images of celebrity faces with varying attributes (e.g., smiling, male, young). These datasets were chosen to capture both object-centric and human-centric image contexts.

For each image, we generated a corresponding textual description using the BLIP model, which employs a Vision Transformer (ViT) and a language decoder to caption visual content. These captions served as the original textual descriptions of the images, serving as a basic description of what is shown in the image.

2.2 Model Architecture

We used the Stable Diffusion v1.5 model for generating images from text prompts. This model works by gradually turning random noise into a coherent image, guided by the meaning of the input caption. It uses a latent diffusion process along with a cross-attention mechanism to align image features with text. Depending on the wording of the prompt, it can produce either realistic or stylized outputs.

For our experiments, we used the available pretrained version from Hugging Face (`CompVis/stable-diffusion-v1-5`). To make sure the model accurately understood each caption, we relied on the default CLIP tokenizer and text encoder (ViT-L/14) built into the pipeline. Images generated from both the original and modified captions were saved for further evaluation.

2.3 Caption Generation

Starting from real-world datasets—specifically CIFAR-10 and CelebA, which contain object-focused and human-centric images respectively—we use the BLIP model to generate captions (original_caption) from raw input images (original_image). These captions are then fed into Stable Diffusion to produce synthetic images (sd_image_original).

2.4 Caption Modification

To test the sensitivity of image generation to language changes, we created two altered versions of each original caption:

Stylistic Captions: Modified to include descriptive phrases such as *"professional photo," "animated version," "art style,"* and *"magazine look."*

Noisy Captions: Injected with unrelated or confusing noise terms like *"random objects," "blurry light," "chaotic texture,"* and *"random street."*

2.5 Image Recaption

To evaluate how well the generated images reflect their textual inputs, we recaption both sd_image_original and sd_image_modified using BLIP (recaption_original and recaption_modified, respectively). We used the ROUGE-L metric to quantify the semantic overlap between the input captions and the corresponding recaptions. The results of these comparisons are as follows:

Original Caption vs. Recaption: This measures how well Stable Diffusion preserves the original semantic content. A relatively high score suggests strong consistency.

Style Caption vs. Recaption: This evaluates how stylistic additions influence meaning retention. The lower score indicates that the semantic representation may have shifted.

Noisy Caption vs. Recaption: This measures the impact of semantically unrelated words.

Original Caption vs. Recaption from Style Prompt: This reveals how much the recaptioned output deviates from the original content.

Original Caption vs. Recaption from Noisy Prompt: This shows how noisy prompts affect the model's ability to return to the original meaning after generation.

2.6 Evaluation Metrics

To gain a comprehensive view of the outputs, we analyzed semantic, perceptual, and visual fidelity using ROUGE-L, CLIP similarity, and Inception Score.

ROUGE-L: It measures the semantic similarity between text sequences and was used to evaluate how closely recaptions matched the original input captions.

CLIP Similarity: It measures how well an image matches a given text by projecting both into a shared embedding space.

Inception Score : The Inception Score estimates the visual quality and diversity of generated images based on how confidently a pretrained classifier can assign labels. Higher value suggests clearer outputs.

3 Experiment

We evaluated our pipeline using three main metrics: ROUGE-L for semantic consistency, CLIP similarity for image-text alignment, and Inception Score for image quality. Comparisons were made across three caption conditions—original, style-modified, and noise-modified—on both CIFAR-10 and CelebA datasets.

3.1 ROUGE-L Score

To evaluate how well semantic content is kept throughout the process, we used ROUGE-L scores to compare input captions with their corresponding recaptions. Higher ROUGE-L values indicate greater overlap in structure and meaning.

Variant	Dataset	Rouge Score
Original vs Recaption Original	CIFAR-10	0.6737
Original vs Recaption Style	CIFAR-10	0.5720
Original vs Recaption Noisy	CIFAR-10	0.5346
Style vs Recaption Style	CIFAR-10	0.6163
Noisy vs Recaption Noisy	CIFAR-10	0.5961
Original vs Recaption Original	CelebA	0.6041
Original vs Recaption Style	CelebA	0.5501
Original vs Recaption Noisy	CelebA	0.4932
Style vs Recaption Style	CelebA	0.6103
Noisy vs Recaption Noisy	CelebA	0.5514

Table 1: ROUGE-L scores across different caption comparison pairs.

CIFAR-10 Results

The highest ROUGE-L score was observed in the Original vs. Recaption Original comparison (0.6737), indicating best maintained semantic structure when the prompt was unaltered. Both style-modified (0.5720) and noise-modified (0.5346) captions led to reduced alignment with the

resulting recaptions, confirming that prompt alteration decrease semantic stability. Comparing style vs. recaption style (0.6163) and noisy vs. recaption noisy (0.5961), the scores suggest that while meaning shifts happen, a fair amount of the semantics is kept within modified context.

CelebA Results

Compared to CIFAR-10, the ROUGE-L scores for CelebA were consistently lower, reflecting the increased sensitivity of human face generation to prompt alterations. The Original vs. Recaption Original scored 0.6041, while style and noise variants decreased to 0.5501 and 0.4932 respectively. Interestingly Style vs. Recaption Style (0.6103) outperformed Original vs. Recaption Original, suggesting that stylistic prompts formed a more consistent loop when evaluated within their own context.

3.2 CLIP Cosine Similarity

CLIP scores measured how closely the generated images aligned with their respective prompts:

Caption Variant	Dataset	CLIP Score
Original	CIFAR-10	0.1661
Style	CIFAR-10	0.1743
Noisy	CIFAR-10	0.1668
Original	CelebA	0.1391
Style	CelebA	0.1385
Noisy	CelebA	0.1301

Table 2: CLIP cosine similarity scores between generated images and prompts across datasets.

CIFAR-10 Results

Interestingly, the highest score was found for style-modified prompts, suggesting that aesthetic cues encouraged the model to generate visually clearer or more distinctive content. Noisy prompts still able to maintain comparable alignment to original prompts.

CelebA Results

The decline in score from original to noise-modified shows the increase sensitivity of facial imagery. Small changes in caption phrasing can result in reduced image–text alignment. Noisy prompts led to the lowest CLIP scores, suggesting that unrelated terms decrease the relevance between the generated face and its intended description.

3.3 Inception Score

Inception Score assessed the visual quality and class diversity of generated images.

Variant	Dataset	Inception Score
Original	CIFAR-10	13.00
Style	CIFAR-10	10.99
Noisy	CIFAR-10	12.43
Original	CelebA	6.39
Style	CelebA	6.13
Noisy	CelebA	5.61

Table 3: Inception Scores across prompt types and datasets.

CIFAR-10 Results

In CIFAR-10, the highest IS was achieved with original captions (13.00), indicating clear outputs when the prompt was unaltered. Style-modified captions dropped to 10.99, reflecting a slight decrease in image coherence. These results suggest that for object-centric images, Stable Diffusion is fairly robust to prompt modifications.

CelebA Results

In contrast, the CelebA results show lower score. This trend mirrors earlier findings from ROUGE and CLIP that facial imagery is more sensitive to prompt language,. The drop under noisy conditions was notable, suggestint that ambiguous text cues disrupt generation in more human-centered domains.

3.4 Drifting Examples: CIFAR-10 and CelebA

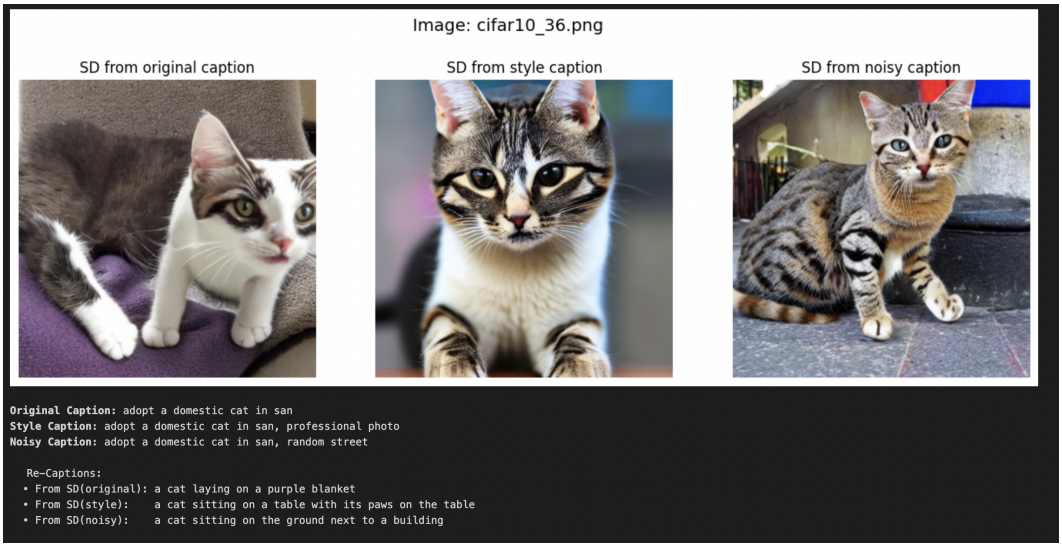

Figure 1: Worst drifting example in CIFAR-10

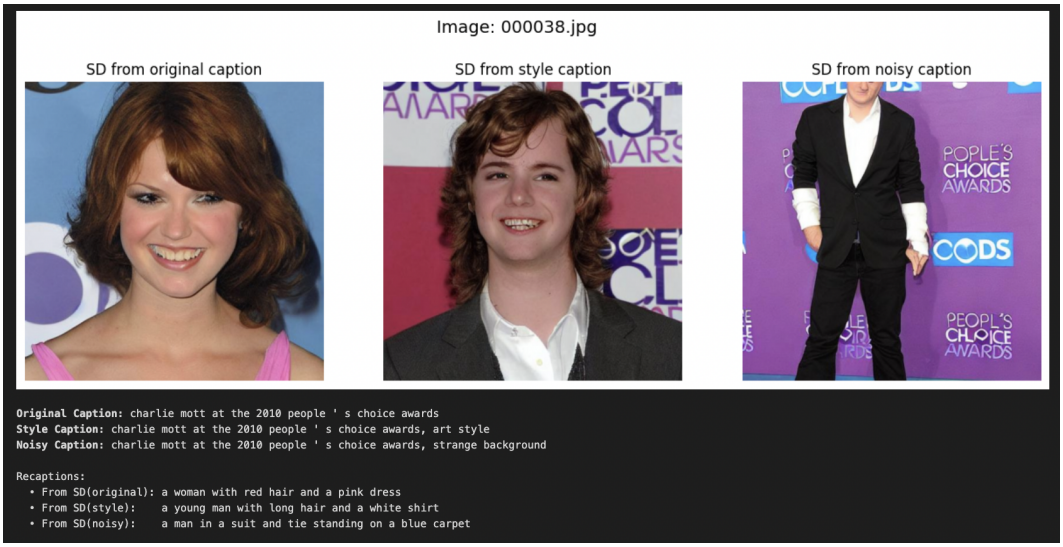

Figure 2: Worst drifting example in CelebA

To further illustrate the effects of caption modifications on image generation and semantic preservation, we present two worst-case examples of style and semantic drift, identified through a combination of metrics.

Figure 1 shows a CIFAR-10 example featuring a domestic cat. While the original caption ("adopt a domestic cat in san") produces a realistic image of a cat lying on a blanket, the style-modified caption

("professional photo") shifts the pose and composition significantly. The noisy caption introduces semantic confusion ("random street"), resulting in a background and posture more detached from the original image. The recaptions confirm this drift, ranging from a cat on a blanket to one sitting on the ground near a building.

Figure 2 highlights a CelebA example involving a celebrity at an award event. The original image shows a woman in a pink dress, aligned with the original caption. The style-modified prompt emphasizes "art style," resulting in a male-presenting individual, while the noisy caption leads to a male figure in a suit on a blue carpet which is more misaligned with the original intent and setting.

These examples demonstrate the sensitivity of text-to-image models like Stable Diffusion to subtle prompt variations and the need for semantic control in generative process.

4 Conclusion

In this project, we explored how changes in caption phrasing influence Stable Diffusion's ability to generate images that reflect intended meaning and style. Using BLIP to caption real images from CIFAR-10 and CelebA, we built a pipeline to generate images from both original and modified captions, then recaptioned the outputs to assess semantic consistency. Our caption modifications included stylistic changes and injected noise to test the model's sensitivity to language variations.

Our findings show difference in how the model handles object-centric versus human-centric imagery. CIFAR-10 images, which include well-defined object categories, showed better stability and semantic consistency across the generation and recaptioning process. Images from the CelebA dataset, which focuses on human faces, were more sensitive to subtle changes in caption phrasing. Some small additions or alterations to captions could lead to noticeable shifts in both the visual outputs and their corresponding recaptions. Stylistic edits in CIFAR-10 typically produced refined or exaggerated versions of original objects, while noisy captions caused minor semantic drift but maintained overall content. For CelebA, stylistic prompts led to aesthetic shifts, and noisy captions sometimes caused facial disruptions or ambiguity. CLIP similarity scores supported these observations, with greater misalignment appearing when noise was introduced, especially for CelebA. While noisy captions consistently degraded performance in both datasets, stylistic modifications had more subtle effects, mostly shifting tone or visual style without fully altering context.

Overall, the results from both datasets highlight the importance of prompt design in text-to-image workflows. Language control is important, especially in domains like facial imagery. Our work offers a framework for analyzing semantic consistency and shows how prompt precision directly impacts visual images generation. Future work could extend this analysis by introducing more diverse datasets and caption edits to better understand how meaning evolves across generation cycles.

5 References

[1] Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.* Proceedings of the 39th International Conference on Machine Learning (ICML), `https://github.com/salesforce/BLIP`

[2] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). *Deep learning face attributes in the wild*. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 3730–3738. `http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html`

[3] Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images* [Technical report]. University of Toronto. `https://www.cs.toronto.edu/~kriz/cifar.html`

[4] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, *29*.