Text-to-Image Quality Evaluation Using Computer Vision Metrics

Author: Ishan Srivastava
Course: Computer Vision
Date: December 2025

1. Introduction

Text-to-image generation has emerged as one of the most exciting developments in artificial intelligence over the past few years. Models like Stable Diffusion, DALL-E, and Midjourney have transformed how we think about visual content creation. These systems can generate remarkably realistic and creative images from simple text descriptions, opening up new possibilities for artists, designers, and content creators.

However, despite the impressive visual results these models produce, evaluating their performance remains a significant challenge. Most assessments of text-to-image models rely heavily on subjective human judgment. A person looks at a generated image and decides whether it looks good or matches the prompt. While human evaluation captures important aspects of image quality, it lacks consistency and reproducibility. Different people may have different opinions about the same image, and the same person might judge differently on different days.

This limitation creates a real problem for researchers and developers working to improve these models. Without objective metrics, it becomes difficult to compare different models, track improvements over time, or identify specific weaknesses that need attention. The field needs standardized, quantitative methods for evaluation.

This project addresses this gap by developing and implementing a comprehensive evaluation framework for text-to-image models. The framework uses established computer vision metrics to provide objective, reproducible assessments of generated image quality. Specifically, I evaluate Stable Diffusion v1.5 across 470 text prompts spanning five distinct categories: objects, attributes, spatial relations, counting, and artistic styles.

The evaluation employs multiple complementary metrics. CLIP Score measures how well generated images align with their text descriptions. Inception Score assesses image quality and diversity. Fréchet Inception Distance compares the distribution of generated images to real photographs. Traditional OpenCV metrics capture low-level properties like brightness, contrast, and sharpness. Together, these metrics provide a multi-dimensional view of model performance that no single measure could offer alone.

2. Literature Review

2.1 CLIP Score

CLIP, which stands for Contrastive Language-Image Pre-training, was introduced by Radford and colleagues at OpenAI in 2021. The model learns to understand the relationship between images and text by training on 400 million image-text pairs collected from the internet. During training, CLIP learns to map both images and text into a shared 512-dimensional embedding space where related concepts cluster together.

For evaluating text-to-image models, CLIP Score measures the cosine similarity between the embedding of a text prompt and the embedding of its corresponding generated image. A higher score indicates that the image and text are semantically closer in the learned embedding space, suggesting better alignment between what was requested and what was generated.

CLIP has become the standard metric for text-to-image evaluation because it correlates well with human judgments of text-image correspondence. Research has shown that when humans rate how well an image matches a description, their ratings tend to agree with CLIP scores. However, CLIP has limitations. It can assign high scores to images that contain the right objects but arrange them incorrectly. It may also be biased toward visual patterns that were common in its training data.

2.2 Inception Score

The Inception Score was proposed by Salimans and colleagues in 2016 as a way to evaluate generative adversarial networks. The metric uses a pre-trained Inception-V3 network, which was originally designed to classify images into 1000 ImageNet categories.

Inception Score measures two complementary properties. First, it assesses image quality by checking whether the classifier can make confident predictions about generated images. If an image is clear and recognizable, the classifier should be confident about what it depicts. Second, it measures diversity by checking whether the generated images span many different categories. A model that only generates one type of image would score poorly on diversity.

Mathematically, Inception Score is calculated as the exponential of the expected KL divergence between the conditional class distribution for individual images and the marginal class distribution across all images. Scores typically range from 1 (poor) to around 50 for the best models, though most good text-to-image models score between 5 and 15.

The main limitation of Inception Score is that it does not compare generated images to any reference set of real images. A model could theoretically achieve a high Inception Score by generating clear, diverse images that look nothing like real photographs.

2.3 Fréchet Inception Distance

Fréchet Inception Distance, or FID, was introduced by Heusel and colleagues in 2017 to address the limitations of Inception Score. Unlike IS, FID explicitly compares generated images to a reference set of real images.

FID works by extracting features from both generated and real images using the Inception-V3 network. It then models each set of features as a multivariate Gaussian distribution, characterized by a mean vector and covariance matrix. The Fréchet distance between these two Gaussian distributions gives the FID score.

Lower FID scores indicate that the distribution of generated images is more similar to the distribution of real images. Scores below 30 are generally considered excellent, while scores above 100 suggest significant differences between generated and real image distributions.

FID is sensitive to several factors including sample size, image preprocessing, and the choice of reference dataset. Research has shown that FID requires at least several hundred images to produce stable estimates. Using too few images can result in unreliable scores.

2.4 Traditional Computer Vision Metrics

Beyond deep learning metrics, traditional computer vision measures provide interpretable information about low-level image properties. Brightness measures the average pixel intensity, indicating whether images tend to be light or dark. Contrast, calculated as the standard deviation of pixel values, shows how much variation exists between light and dark regions. Saturation measures color vibrancy in the HSV color space. Edge density, computed using Canny edge detection, indicates image sharpness and level of detail.

These metrics complement the semantic measures from deep learning. While CLIP tells us whether the right objects are present, OpenCV metrics tell us about the visual characteristics of how those objects are rendered.

3. Data and Framework

3.1 Dataset Construction

A critical part of this project was designing a diverse set of text prompts that would test different aspects of the model's capabilities. I created 470 prompts organized into five categories, each targeting a specific skill.

The Objects category contains 100 simple prompts describing single objects like "a red bicycle" or "a coffee mug." These test the model's basic ability to recognize and render common items. The Attributes category includes 90 prompts that add descriptive modifiers, such as "a shiny

ceramic mug" or "a rusty metal bicycle." These test whether the model understands adjectives and can apply visual properties correctly.

The Relations category contains 90 prompts involving spatial relationships between objects, like "a cat sitting on a sofa" or "a book on a wooden table." Spatial reasoning is known to be challenging for text-to-image models, so this category tests a potential weakness. The Counting category has 100 prompts specifying exact numbers, such as "three apples on a plate" or "five birds on a wire." Counting is another documented weakness of diffusion models.

Finally, the Styles category includes 90 prompts requesting specific artistic styles, like "a watercolor painting of a mountain" or "a pencil sketch of a portrait." These test the model's ability to generate images in different visual styles beyond photorealism.

3.2 Reference Dataset

For computing FID, I needed a reference set of real images. I used 1000 images from the COCO (Common Objects in Context) validation set, downloaded through the HuggingFace datasets library. COCO is a standard benchmark dataset containing photographs of everyday scenes with common objects.

Using 1000 reference images ensures statistical stability in the FID calculation. Initial attempts with fewer images produced unreliable results, highlighting the importance of adequate sample size for distributional metrics.

3.3 Technical Framework

All experiments were conducted on Google Colab using a Tesla T4 GPU with 16GB of video memory. Google Drive provided persistent storage for generated images, results, and intermediate files. This setup offers free access to GPU computing resources suitable for running Stable Diffusion.

For image generation, I used Stable Diffusion v1.5 accessed through the HuggingFace Diffusers library. Images were generated at 512x512 resolution using 30 inference steps and a guidance scale of 7.5. Each prompt was assigned a deterministic seed based on its ID to ensure reproducibility.

For evaluation, CLIP scores were computed using the openai/clip-vit-base-patch32 model. Inception Score and FID used features extracted from Inception-V3. OpenCV provided functions for computing traditional image metrics.

4. System Design

4.1 Pipeline Overview

The evaluation system follows a sequential pipeline architecture. Text prompts flow into the image generation stage, which produces synthetic images. These images then undergo feature extraction using multiple models. Finally, various metrics are computed and aggregated for analysis.

This modular design allows each component to be developed and tested independently. If a problem is found in one stage, it can be fixed without affecting other parts of the system.

4.2 Phase-wise Architecture

The implementation is organized into six distinct phases, each handling a specific part of the pipeline.

Phase 1 handles project setup. This includes mounting Google Drive for persistent storage, installing required Python packages, and creating the prompt dataset. The prompts are saved to a CSV file for easy loading in subsequent phases.

Phase 2 performs image generation. The Stable Diffusion model is loaded onto the GPU, and images are generated for each prompt. Progress is tracked and displayed, and images are saved incrementally to prevent loss if the process is interrupted. Memory management through periodic cache clearing prevents out-of-memory errors.

Phase 3 applies OpenCV preprocessing and extracts traditional metrics. Each image is loaded and analyzed to compute brightness, contrast, edge density, and saturation. Results are aggregated by category to reveal patterns.

Phase 4 computes CLIP scores. Both prompts and images are encoded into the CLIP embedding space, and cosine similarity is calculated for each pair. This phase required careful implementation to ensure correct normalization.

Phase 5 handles Inception Score and FID computation. Reference images are downloaded from COCO, features are extracted from both generated and reference images using Inception-V3, and the metrics are calculated. FID uses the standard formula comparing Gaussian distributions of features.

Phase 6 performs final analysis and visualization. Results from all metrics are combined, statistical summaries are computed, and visualizations are generated to communicate findings.

5. Implementation Process

5.1 Initial Implementation

The first version of this system was developed for the midterm assessment. That implementation used 100 prompts with 20 per category. While the basic pipeline worked, the results revealed significant problems that needed correction.

The initial CLIP score came out to 0.9551, which seemed excellent at first glance. However, this value is far higher than what published research reports for Stable Diffusion, which typically falls between 0.25 and 0.35. Such an extreme discrepancy indicated a bug rather than exceptional performance.

The initial FID score was 429.45, which is very poor. Investigation revealed that the reference image download had mostly failed, leaving only 5 images instead of the intended 500. FID computed with such a small reference set is statistically meaningless.

The Inception Score of 6.88 was the only metric in a reasonable range, suggesting that the feature extraction and IS calculation were implemented correctly.

5.2 Problems Identified

The CLIP score error stemmed from incorrect normalization. The original code used a sigmoid function to normalize the raw similarity scores:

```
batch_scores_normalized = 1 / (1 + np.exp(-batch_scores / 10)
```

This transformation compressed scores into the 0-1 range but did not produce proper cosine similarities. The correct approach is to normalize the embedding vectors to unit length before computing their dot product.

The FID reference image problem occurred because the direct download URLs for COCO images were unreliable. Many requests failed silently, and the code did not adequately check for successful downloads before proceeding.

The small sample size of 100 images, while sufficient for basic testing, limited the statistical power of the analysis. More prompts per category would provide more reliable estimates of category-specific performance.

5.3 Improvements Made

The CLIP computation was corrected to use proper cosine similarity:

```
image_embeds = image_embeds / image_embeds.norm(dim=-1, keepdim=True)
text_embeds = text_embeds / text_embeds.norm(dim=-1, keepdim=True)
similarities = (image_embeds * text_embeds).sum(dim=-1)
```

This normalizes both embedding vectors to unit length, then computes their dot product. The result is a true cosine similarity bounded between -1 and 1, with typical values for Stable Diffusion falling in the 0.25-0.35 range.

For reference images, I switched to the HuggingFace datasets API, which provides reliable streaming access to COCO. This successfully downloaded 1000 reference images without failures.

The prompt dataset was expanded to 470 prompts with roughly 100 per category. This provides much more robust statistics and allows meaningful comparisons between categories.

5.4 Deployment Attempt

Beyond the core evaluation system, I attempted to create a web application that would let users analyze their own prompts. The Prompt Quality Analyzer was deployed on HuggingFace Spaces and is accessible at huggingface.co/spaces/Ishansri13/prompt-quality-analyzer.

The current version analyzes prompt structure and estimates likely quality based on the presence of descriptive words, style specifications, and other elements learned from the evaluation data. However, actual image generation through the free API proved unreliable due to rate limits and availability issues. A fully functional version would require paid GPU resources.

6. System Performance and Results

6.1 Final Metrics Comparison

The improvements made between midterm and final submission dramatically changed the results:

Metric	Midterm	Final	Expected Range
CLIP Score	0.9551	0.2891	0.25-0.35
Inception Score	6.88	16.49	5-15
FID Score	429.45	156.36	50-150

The corrected CLIP score of 0.2891 falls squarely within the expected range for Stable Diffusion v1.5, validating that the implementation now works correctly. The Inception Score increased from 6.88 to 16.49, reflecting both the larger dataset and greater diversity of prompts. The FID score dropped from the meaningless 429 to 156, which is interpretable though higher than typical benchmarks.

6.2 CLIP Score Analysis

The mean CLIP score across all 470 images was 0.2891 with a standard deviation of 0.0504. Scores ranged from 0.0971 to 0.3837, showing substantial variation in how well different prompts were satisfied.

Breaking down by category reveals interesting patterns. Objects achieved the highest mean score at 0.300, confirming that simple, concrete prompts produce the best alignment. Styles came second at 0.294, showing the model handles artistic style requests well. Attributes scored 0.289, Relations scored 0.284, and Counting came last at 0.279.

The relatively poor performance on Counting confirms a known limitation of diffusion models. When asked to generate a specific number of objects, the model frequently produces the wrong count. The Relations category also struggled, suggesting spatial reasoning remains challenging.

6.3 Inception Score Analysis

The Inception Score of 16.49 with standard deviation 3.07 indicates excellent image quality and diversity. This score exceeds typical reported values for Stable Diffusion, likely because the diverse prompt categories encourage varied outputs spanning many visual styles and object types.

The improvement from 6.88 at midterm to 16.49 in the final version primarily reflects the expanded dataset. With more prompts covering more categories, the generated images naturally span a wider range of visual concepts, boosting the diversity component of IS.

6.4 FID Score Analysis

The FID score of 156.36 is higher than the 25-50 range typically reported for Stable Diffusion. However, this does not indicate poor performance. Rather, it reflects an intentional domain gap between my generated images and the COCO reference set.

My prompts include many artistic styles like watercolor paintings, pixel art, and pencil sketches. These intentionally differ from natural photographs. The COCO dataset contains only real photographs. Comparing artistic renderings to photographs naturally produces a large distributional difference.

This finding illustrates an important principle: FID scores are only meaningful when comparing similar types of images. Using FID to compare artistic generations to photographs conflates style differences with quality differences.

6.5 OpenCV Metrics Analysis

The traditional metrics reveal consistent visual characteristics across generated images. Mean brightness of 116.52 on a 0-255 scale indicates balanced exposure, neither too dark nor too bright. Contrast of 54.99 shows healthy dynamic range. Saturation of 81.33 indicates vibrant colors. Edge density of 0.087 suggests moderate detail levels.

Comparing across categories, Objects showed the highest contrast at 58.63, likely because isolated objects against backgrounds create clear edges. The Counting category had the highest saturation at 89.94, an interesting finding that might reflect the model compensating for counting difficulty with enhanced colors.

6.6 Correlation Analysis

Examining correlations between CLIP score and OpenCV metrics reveals that semantic alignment is largely independent of low-level image properties. The strongest correlation was between CLIP and contrast at -0.30, suggesting the model may slightly prioritize semantic correctness over high contrast when tradeoffs exist.

Other correlations were weak, all below 0.15 in absolute value. This independence is actually desirable. It means the model can achieve good text alignment across a range of visual styles without being constrained to particular brightness or saturation levels.

6.7 Strengths of the System

The evaluation framework successfully provides objective, reproducible measurements of text-to-image quality. The multi-metric approach captures different quality dimensions that single metrics would miss. Category-based analysis reveals specific model weaknesses that aggregate scores would obscure. Fixed random seeds ensure that results can be exactly reproduced.

6.8 Weaknesses of the System

The FID score is not directly comparable to published benchmarks due to the domain gap between artistic prompts and photographic references. The evaluation covers only one model, Stable Diffusion v1.5, without comparisons to other systems. There is no human evaluation to validate that the metrics actually correspond to perceived quality.

7. Conclusion and Future Directions

7.1 Summary

This project successfully developed and implemented a comprehensive evaluation framework for text-to-image generative models. Through careful debugging and iterative improvement, critical errors in the initial implementation were identified and corrected. The final system produces realistic, interpretable metrics that align with published research findings.

The key results demonstrate that Stable Diffusion v1.5 achieves good text-image alignment with a CLIP score of 0.289 and produces high-quality, diverse images as shown by the Inception Score of 16.49. The analysis also confirmed known model limitations, particularly in counting and spatial reasoning tasks.

7.2 Key Lessons Learned

This project taught several important lessons about implementing machine learning evaluation systems. First, always verify metric computations against expected ranges from literature. My initial CLIP score of 0.95 should have immediately raised red flags since it far exceeded published values.

Second, sample size critically affects distributional metrics like FID. The meaningless initial FID resulted directly from having only 5 reference images. Planning for adequate data collection should happen before running experiments.

Third, domain gaps between test and reference data affect distributional comparisons. FID measured against photographs cannot directly assess the quality of artistic renderings. Metric interpretation must account for what is actually being compared.

Fourth, multiple metrics provide complementary insights. CLIP captures semantic alignment, IS captures quality and diversity, FID captures distributional similarity, and OpenCV metrics capture visual properties. No single metric tells the whole story.

7.3 Future Improvements

Several directions could extend and improve this work. Comparing multiple models including SD-XL, DALL-E 3, and Midjourney would provide valuable benchmarking data. Adding human evaluation studies would validate whether the metrics actually predict perceived quality.
More fine-grained analysis could separately assess counting accuracy, color correctness, and spatial arrangement accuracy. A larger dataset with 1000 or more prompts would improve statistical power. Adding LPIPS (Learned Perceptual Image Patch Similarity) would provide another perspective on perceptual quality.

On the deployment side, securing GPU resources would enable real image generation in the web application, making the tool more practically useful.

7.4 Practical Applications

The evaluation framework developed here has several practical applications. It enables data-driven model selection when choosing which text-to-image system to deploy. It identifies specific weaknesses requiring improvement, guiding development priorities. It provides standardized benchmarks for comparing research contributions. And it enables automated quality assurance testing in production systems.

As text-to-image generation continues advancing rapidly, objective evaluation methods will become increasingly important for measuring progress and ensuring quality. This project contributes a working implementation of such methods along with lessons learned from the development process.

References

1. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision. International Conference on Machine Learning (ICML).*

2. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). *GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems (NeurIPS).*

3. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). *Improved techniques for training GANs. Advances in Neural Information Processing Systems (NeurIPS).*

4. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft COCO: *Common objects in context. European Conference on Computer Vision (ECCV).*

5. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2022). *CLIPScore: A reference-free evaluation metric for image captioning. arXiv preprint.*
6.
7. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-resolution image synthesis with latent diffusion models. Conference on Computer Vision and Pattern Recognition (CVPR).*
8. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., & Wang, O. (2018). *The unreasonable effectiveness of deep features as a perceptual metric. Conference on Computer Vision and Pattern Recognition (CVPR).*