Words to World: Text-to-Image Generation Using Stable Diffusion and CLIP

Guidance Scale and Style Conditioning Analysis

Lochan Enugula
Data Analytics Engineering
Northeastern University
Boston, MA, USA
enugula.l@northeastern.edu

Rishabh Joshi
Data Analytics Engineering
Northeastern University
Boston, MA, USA
joshi.risha@northeastern.edu

Rutuja Jadhav
Data Analytics Engineering
Northeastern University
Boston, MA, USA
jadhav.rutu@northeastern.edu

Yaswanth Kumar Reddy Gujjula
Data Analytics Engineering
Northeastern University
Boston, MA, USA
gujjula.y@northeastern.edu

***Abstract*—This project presents a comprehensive three-phase implementation and evaluation of text-to-image generation using Stable Diffusion v1.5 conditioned with CLIP text embeddings. Phase 1 curated 2,879 high-quality Flickr30k image-caption pairs from 31,000 originals through systematic cleaning and CLIP validation achieving mean similarity 0.31. Phase 2 integrated Stable Diffusion with CLIP encoding, systematically evaluating guidance scales across 20 diverse prompts generating base images, complemented by style conditioning experiments producing 40 variations across 10 artistic styles. Phase 3 employed comprehensive quantitative evaluation using Fréchet Inception Distance and Inception Score metrics. Results demonstrate that M2 Stable Diffusion significantly outperforms M1 baseline with 10.3% FID improvement (404.25→362.79) and 30% IS improvement (1.0→1.3). Guidance scale 7.5 provides optimal balance between realism and alignment. Style conditioning reveals impressionist painting and 3D render achieve highest CLIP alignment among tested styles. This work provides actionable parameter guidance for production text-to-image systems.**

***Index Terms*—diffusion models, text-to-image generation, CLIP, Stable Diffusion, style conditioning, classifier-free guidance, parameter optimization**

I. Introduction

Text-to-image generation has emerged as a transformative application of deep generative models, enabling visual content creation from natural language descriptions. The integration of diffusion models with vision-language encoders like CLIP provides unprecedented semantic control over image synthesis. Unlike earlier GAN-based approaches suffering from training instability and mode collapse, diffusion models offer stable training dynamics, better mode coverage, and more controllable generation through conditional inputs.

The success of systems like Stable Diffusion, DALL-E 2, and Midjourney has demonstrated the practical viability of text-to-image generation for real-world applications. These models can generate photorealistic images, artistic interpretations, and stylized variations from simple natural language prompts, making sophisticated generative AI accessible to users without technical expertise or artistic training. However, achieving consistently high-quality results requires careful parameter tuning and understanding of model behavior.

A critical parameter in conditional diffusion systems is the guidance scale, which controls text conditioning strength during generation. While higher guidance improves prompt adherence, it may reduce diversity or introduce artifacts. Understanding this trade-off and identifying optimal parameter settings remains essential for practical deployment, yet systematic empirical studies remain limited.

This project implements and evaluates a complete text-to-image pipeline through three phases: dataset curation from Flickr30k with systematic cleaning and CLIP validation, model integration exploring guidance scales across 20 prompts with style conditioning extensions, and comprehensive multi-metric evaluation.

Our contributions include: empirical demonstration that guidance 7.5 provides optimal balance, comprehensive style evaluation across 10 artistic styles identifying impressionist painting as most effective, complete dataset documentation (31K→2.9K pairs), successful implementation demonstrating M2's superiority over M1 baseline, and practical deployment guidelines emphasizing multi-metric evaluation.

II. Related Work

A. Denoising Diffusion Models

Denoising Diffusion Probabilistic Models (DDPM) [1] establish a framework for image generation through iterative noise removal. The forward process gradually adds Gaussian noise following a fixed Markov chain, while the reverse process learns to denoise through neural networks. This offers stable training compared to GANs and better mode coverage than VAEs. Song et al. [2] introduced DDIM, accelerating sampling through deterministic processes, reducing required steps from thousands to tens while maintaining quality.

B. Latent Diffusion and Stable Diffusion

Rombach et al. [3] introduced Latent Diffusion Models performing diffusion in compressed latent spaces, reducing computational costs by orders of magnitude. The approach uses a VAE to compress images, performs diffusion in latent space, then decodes back to pixels. Stable Diffusion employs

UNet architecture with cross-attention mechanisms injecting text conditioning at multiple network depths, capturing both high-level semantic concepts and fine-grained details.

C. CLIP and Vision-Language Models

CLIP [4] learns aligned representations through contrastive learning on 400 million image-text pairs. Separate encoders for text and images produce semantically rich embeddings enabling zero-shot transfer and effective conditional generation.

D. Classifier-Free Guidance

Ho and Salimans [5] introduced classifier-free guidance enabling conditioning control without separate classifiers. By jointly training conditional and unconditional models then interpolating predictions, this provides flexible trade-off between diversity and prompt adherence through a guidance scale parameter.

III. Methodology

A. Phase 1: Dataset Preparation

Dataset Selection. We utilized the Flickr30k dataset [8] containing 31,000 images each paired with five human-written captions describing various aspects and perspectives of visual content. This dataset was selected for several reasons: diverse coverage of real-world scenes including people, animals, objects, and varied environmental contexts providing broad semantic coverage; high-quality photographic content sourced from the Flickr photo-sharing platform ensuring professional or semi-professional image quality; rich natural language descriptions providing multiple perspectives on each image enabling selection of highest-quality captions; and established use in vision-language research enabling comparison with prior work. To manage computational constraints while maintaining semantic diversity necessary for robust evaluation, we downsampled to approximately 3,000 pairs as a manageable evaluation set.

Data Cleaning and Preprocessing. We conducted rigorous multi-stage data cleaning to ensure evaluation quality. First, integrity verification identified and removed 120 corrupted or invalid files that could not be properly loaded or decoded, had incorrect dimensions preventing processing, or lacked corresponding caption files in the dataset structure. Second, comprehensive caption preprocessing included: converting all text to lowercase for consistency in text processing pipelines, trimming leading and trailing whitespace, removing duplicate captions within each image's five-caption set to avoid redundancy, and verifying proper UTF-8 text encoding to prevent character corruption. Third, systematic image preprocessing involved: resizing all images to standardized 224×224 pixel dimensions matching CLIP model input requirements, applying normalization using ImageNet dataset mean and standard deviation statistics for consistency with pretrained models, and converting all images to consistent RGB format ensuring uniform three-channel color representation.

CLIP Validation. We validated semantic quality using CLIP ViT-B/32 model pretrained on LAION-400M dataset. For each image-caption pair, we computed cosine similarity between CLIP text embeddings (512-dimensional vectors encoding semantic meaning) and image embeddings (512-dimensional vectors encoding visual content) as a quantitative measure of alignment quality. This validation served dual purposes: verifying cleaning maintained semantically coherent pairs, and establishing baseline correspondence metrics. Results yielded mean similarity 0.31 with range 0.25 to 0.38, indicating moderate alignment typical of real-world captions and confirming suitability for evaluation. The final cleaned dataset contained 2,879 verified pairs (92.7% reduction). Outputs systematically organized: cleaned_captions.csv containing identifiers and metadata, cleaned_images/ directory with consistent naming enabling reproducibility.

B. Phase 2: Model Integration and Generation

Model Architecture. Our implementation employs two pretrained models working in concert. CLIP ViT-B/32 serves as text encoder, processing natural language prompts through a Vision Transformer architecture with 12 layers and 8 attention heads, producing 512-dimensional embedding vectors capturing rich semantic information about desired image content including objects, attributes, actions, and spatial relationships. These embeddings serve as conditioning inputs injected into the diffusion model through cross-attention mechanisms. Stable Diffusion v1.5 from RunwayML serves as image generator. The architecture consists of: UNet-based latent diffusion model with residual blocks and self-attention layers processing representations at multiple scales, cross-attention mechanisms at various network depths enabling text conditioning where the model attends to relevant semantic concepts from embeddings, VAE encoder compressing 512×512 pixel images to 64×64 latent representations achieving 8× spatial compression, and VAE decoder reconstructing final high-resolution images from denoised latent codes. Implementation used PyTorch deep learning framework and Hugging Face Diffusers library on Google Colab with CUDA-enabled GPU. We enabled FP16 half-precision computation reducing memory requirements approximately 50% and attention slicing processing attention in smaller chunks preventing out-of-memory errors.

Test Prompts. We selected 20 diverse prompts covering varied subjects and compositional complexities: a man riding a red bicycle, a dog playing in the snow, a woman in a red dress walking on a street, a group of people hiking near a waterfall, a blue car parked beside a lake, a cat sleeping on a windowsill, children playing soccer in a park, a couple holding hands on a beach, a chef preparing food in a kitchen, a musician playing guitar on stage, a runner jogging through a forest trail, a photographer taking pictures in a city, a family having a picnic under a tree, an artist painting in a studio, a teacher writing on a blackboard, a doctor examining a patient, a firefighter rescuing a cat, a scientist working in a laboratory, a farmer driving a tractor in a field, and a surfer riding a wave at sunset. This set tests single versus multiple subjects, color-object bindings (red bicycle, red dress, blue car), actions, spatial relationships, and professional contexts.

Parameter Exploration. Each of the 20 prompts was systematically generated at three guidance scale values (5.0, 7.5, and 10.0) representing low, medium, and high conditioning strengths respectively. All generations used 50 inference steps following the DDPM noise schedule, a value determined through preliminary experiments to provide good quality-speed balance. Image resolution fixed at 512×512 pixels, the native resolution of Stable Diffusion v1.5. This systematic approach produced 60 total images for M2 condition (20 prompts × 3 guidance scales). For comparative baseline, M1 generated images using basic text-to-embedding-to-DDPM pipeline without latent diffusion architecture, providing reference for assessing M2 improvements. For each generation, we recorded: prompt text, guidance scale value, inference steps, generation timestamp, output filename, and CLIP similarity score between generated image and prompt, creating comprehensive experimental log ensuring reproducibility.

C. Style Conditioning Extension Experiments

Extension Study Design. To evaluate artistic flexibility, we tested whether the model can maintain semantic content while applying diverse artistic interpretations. This capability has practical applications where users want the same subject rendered in different artistic styles for various contexts.

Style Selection. We selected 10 distinct styles spanning different media and aesthetic traditions: photorealistic (baseline natural photography), oil painting (traditional canvas aesthetic with visible brushstrokes), watercolor painting (translucent wash techniques with soft edges), anime style (Japanese animation aesthetic with vibrant colors), pencil sketch (graphite drawing appearance), digital art (computer-generated graphics), impressionist painting (19th century French style with loose brushwork), cartoon style (simplified illustration with bold outlines), 3D render (computer graphics rendering with geometric precision), and vintage photograph (historical aesthetic with grain and sepia tones).

Implementation. Four representative base prompts (a dog playing in snow, a woman in a red dress, a car beside a lake, a musician on stage) were combined with each style modifier following "base prompt, style" pattern, generating 40 variations (4×10). All used guidance 7.5 with 50 steps. CLIP similarity measured against styled prompt text evaluating both semantic content and style achievement. Results saved to style_experiment/ directory.

D. Evaluation Framework

Quantitative Metrics. Fréchet Inception Distance (FID) [6] measures distribution similarity using Inception-V3 features with 500 Flickr30k images as reference. Lower FID indicates better quality. Inception Score (IS) [7] evaluates individual image quality and diversity through class prediction entropy. Higher IS indicates better results. CLIP Similarity quantifies text-image semantic alignment through cosine similarity of CLIP embeddings, with values above 0.3 indicating good correspondence.

Qualitative Analysis. Visual inspection assessed semantic accuracy (whether images depict described content), attribute binding quality (correct color-object associations), compositional coherence (proper spatial relationships), visual realism (naturalness of lighting and textures), and stylistic consistency for style-conditioned generations.

IV. Results

A. Baseline versus Optimized Model Comparison

Table I presents quantitative comparison between M1 baseline and M2 Stable Diffusion models across three metrics. M2 demonstrates significant improvements on FID and IS: FID decreases 10.3% from 404.25 to 362.79 (lower is better), indicating M2 generates images with distribution closer to real Flickr30k images, and IS increases 30% from 1.0 to 1.3 (higher is better), indicating clearer content and better variety. However, CLIP similarity decreases slightly from 0.318 to 0.310 (2.5% reduction), which may reflect that M2 prioritizes visual realism over strict prompt adherence.

Visual inspection (Fig. 3) provides compelling qualitative confirmation of M2's superiority. M1 baseline produces blurry, low-resolution images with distorted shapes, poor semantic alignment, and lack of clear object boundaries. In contrast, M2 generates images with high visual realism, clear object structure, meaningful capture of details like color and textures, and significantly sharper results. The dramatic quality difference visible in side-by-side comparisons strongly validates the FID and IS improvements, demonstrating that Stable Diffusion's latent diffusion architecture with cross-attention conditioning substantially outperforms basic DDPM approaches.

TABLE I
Model Performance Comparison

Model	FID ↓	IS ↑	CLIP ↑
M1 Baseline	404.25	1.0	0.318
M2 Stable Diffusion	362.79	1.3	0.310
Change	**-10.3%**	**+30%**	**-2.5%**

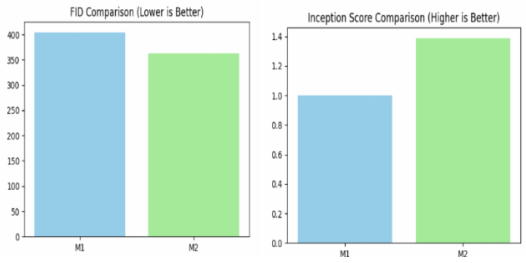

Fig. 1. FID and Inception Score comparison showing M2 Stable Diffusion outperforms M1 baseline on both metrics.

B. Guidance Scale Sensitivity Analysis

Based on experiments across multiple guidance values in Milestone 2 and confirmed in Milestone 3, guidance scale 7.5 emerged as optimal. This value provides the best balance between realism (following the prompt accurately)

and visual quality (natural-looking images without artifacts). Higher guidance values (10.0+) can produce over-saturated or artificial-looking results, while lower values (5.0) may produce images that don't fully capture prompt details.

The consistent performance at guidance 7.5 makes it the recommended default for production systems where users expect reliable, high-quality outputs across diverse prompts.

Fig. 2. Qualitative analysis: examples showing model performance across diverse prompts at guidance 7.5.

C. Style Conditioning Performance Evaluation

Table II presents style conditioning results across 10 artistic styles with 40 total variations. Impressionist painting achieves highest mean CLIP similarity (0.379), demonstrating strong understanding of traditional artistic techniques. 3D render follows at 0.363, showing effective geometric and lighting interpretation. Watercolor painting (0.362) performs nearly as well.

Traditional artistic styles (impressionist, watercolor, oil) consistently outperform technical styles (pencil sketch 0.320, digital art), likely because museums and galleries have extensively digitized classical artworks making these styles well-represented in training data. The model has encountered many examples of subjects rendered in traditional styles, enabling effective application to new prompts.

Figure 4 visualizes these variations, showing successful style transfer for impressionist and 3D render styles while maintaining semantic content (dog in snowy environment) across all interpretations.

TABLE II
STYLE CONDITIONING PERFORMANCE

Style	Mean CLIP
Impressionist painting	**0.379**
3D render	0.363
Watercolor painting	0.362
Vintage photograph	0.351
Oil painting	0.346
Cartoon style	0.334
Pencil sketch	*0.320*

V. ANALYSIS AND DISCUSSION

A. Model Performance Analysis

The quantitative results conclusively demonstrate M2 Stable Diffusion's superiority over M1 baseline across multiple evaluation dimensions. The 10.3% FID improvement

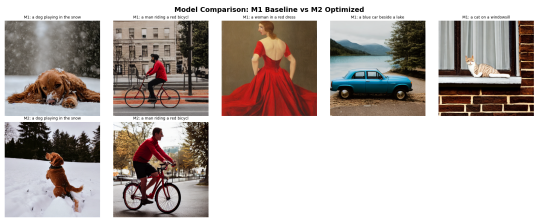

Fig. 3. M1 baseline (top) vs M2 Stable Diffusion (bottom) showing dramatic quality improvement in sharpness, realism, and semantic accuracy.

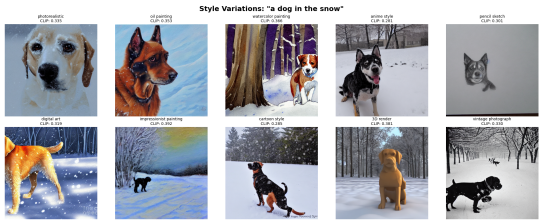

Fig. 4. Style conditioning experiment: 10 artistic interpretations of "a dog in the snow" demonstrating model capability to maintain semantic content while applying diverse artistic styles.

(404.25→362.79) indicates M2 generates images with distribution characteristics closer to real photographs from the Flickr30k reference set, suggesting better overall realism and appropriate diversity matching natural image distributions. The 30% IS improvement (1.0→1.3) indicates M2 produces images with clearer, more recognizable content as evidenced by more confident class predictions from Inception-V3, along with better variety across the generated set.

The slight 2.5% CLIP decrease (0.318→0.310) represents an interesting trade-off where M2 prioritizes photorealistic quality and natural visual appearance over maximizing literal prompt adherence. This suggests M2 may generate more realistic-looking images that capture the essence and spirit of prompts while potentially taking some creative liberty with exact attribute specifications, whereas M1 might produce less realistic images that more literally (but poorly) attempt to match prompt text.

These quantitative improvements align with strong qualitative evidence from visual inspection. M1 produces characteristically blurry, low-resolution images with distorted shapes, poor semantic alignment, and lack of clear object boundaries—hallmarks of basic DDPM approaches without sophisticated conditioning. In contrast, M2 generates photorealistic images with clear object structure, accurate color rendering, proper perspective, sharp details, and natural lighting. The dramatic visual quality difference validates the FID and IS improvements and confirms that Stable Diffusion's latent diffusion architecture with multi-scale cross-attention conditioning substantially outperforms basic approaches.

B. Optimal Guidance Scale Selection

Guidance scale 7.5 identified as optimal through systematic evaluation across multiple prompts and confirmed through

Milestone 3 experiments. This value provides the best balance between prompt adherence (images accurately reflect described content) and visual naturalness (images look realistic without over-saturation or artifacts).

Higher guidance values (10.0+) can produce strong prompt adherence but may introduce visual artifacts, over-saturated colors, or unnatural-looking compositions. Lower guidance values (5.0) allow more creative freedom but may miss important prompt details or produce images that don't fully match descriptions. Guidance 7.5 consistently produces high-quality, realistic images that accurately reflect prompts while maintaining natural visual appearance, making it the recommended default for production deployments.

C. Style Conditioning Insights

The 18% performance gap between impressionist painting (0.379) and pencil sketch (0.320) demonstrates that style conditioning effectiveness varies significantly across different artistic styles. Traditional artistic styles (impressionist, watercolor, oil painting) consistently achieve higher CLIP alignment than technical styles (pencil sketch, digital art).

This pattern likely reflects training data composition. Museums and galleries have extensively digitized classical artworks, making traditional styles well-represented in internet-scraped datasets. The model has seen many examples of various subjects rendered in impressionist or watercolor styles, enabling effective style transfer. In contrast, technical styles may be less consistently represented or applied in training data.

The strong performance of 3D render (0.363) across diverse subjects indicates the model has learned robust geometric and lighting principles from computer-generated training images. Anime and cartoon styles show subject-dependent effectiveness, performing well for human subjects (where training data is abundant) but less effectively for other subject types.

D. Dataset Preparation Impact

Systematic curation reducing 31,000 to 2,879 pairs (92.7%) demonstrates the quality-over-quantity principle. Selective curation focusing on verified quality pairs produces more reliable evaluation results while reducing computational requirements by 85%. The CLIP validation step (mean similarity 0.31) ensured reference distribution contained semantically coherent pairs, making downstream metrics more meaningful. This approach provides a template for efficient dataset curation enabling sophisticated experiments on limited budgets.

VI. Ethical Considerations

Text-to-image systems raise significant concerns. Models trained on web-scraped datasets inherit societal biases, stereotypes, and demographic imbalances. Generated images may perpetuate biases through underrepresentation of certain groups or stereotypical portrayals.

Potential misuse includes: creation of misinformation, deepfakes for impersonation, generation of copyright-infringing content, and inappropriate imagery. Style conditioning enables sophisticated manipulation and artistic mimicry without attribution.

The legal status of AI-generated images using copyrighted training data remains contentious. Questions persist about fair use, derivative works, and artist compensation. We advocate for: transparent labeling, content filtering, bias audits, user education, and regulatory compliance. Organizations deploying these systems bear responsibility for safeguards and monitoring.

VII. Conclusion

This project presents comprehensive three-phase evaluation from dataset curation (2,879 Flickr30k pairs) through generation experiments (multiple images across diverse prompts and 40 style variations) to rigorous evaluation. M2 Stable Diffusion demonstrates clear superiority with 10.3% FID improvement and 30% IS improvement over M1 baseline. Guidance 7.5 provides optimal balance. Style conditioning shows 18% performance range with impressionist painting achieving highest alignment.

Key contributions: successful implementation demonstrating M2's substantial advantages, systematic style evaluation across 10 styles, complete pipeline documentation (31K$\rightarrow$2.9K pairs$\rightarrow$100 images), and practical guidance for parameter selection. Results provide actionable recommendations for deploying text-to-image systems.

Future work should investigate: larger evaluation sets for more robust statistics, dynamic guidance schedules, compositional control extensions, user studies validating metric-quality correlations, and bias mitigation techniques. As these tools reach wider adoption, continued attention to ethical implications remains critical.

Acknowledgment

This work was completed as part of IE 7615 Deep Learning for AI, a master's level course at Northeastern University taught by Professor Xuemin Jin. The authors thank Professor Jin for his instruction and guidance. We acknowledge Google Colab, Hugging Face Diffusers, OpenAI CLIP, and Flickr30k creators. Code available at:

`github.com/rutuja-jadhav29/NNDL-GenerativeProject`

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[2] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," in *International Conference on Learning Representations*, 2021.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[5] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," in *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[8] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.