Comparative Analysis of Stable Diffusion and Conditional DC-GANs: A Multi-Dataset Evaluation Framework

Ansh Bhatnagar
Jeronimo Adames-Baena

June 16, 2025

Abstract

This study investigates and compares the performance of two prominent image generation models, Stable Diffusion (SD) and Conditional Deep Convolutional Generative Adversarial Networks (cDC-GANs), in synthesizing images with fine-grained facial attributes. The CelebA dataset, widely used in facial attribute recognition, serves as the benchmark for the first half of our project. The image generation task was restricted to four categories: woman with bangs, woman without bangs, man with bangs, and man without bangs. Stable Diffusion, as a zero-shot generative model guided by engineered prompts, was evaluated against three different variants of cDC-GANs trained with different hyperparameter settings and loss functions.

The other half of our project was based on the CIFAR-10 dataset, which was relabeled to focus on three classes: airplanes, automobiles, and trucks. As we did with the CelebA dataset, we trained four variants of cDC-GANs, each using a unique combination of optimizers (Adam or RMSprop) and loss functions (BCEWithLogits or Hinge), to evaluate their generative capacity.

Evaluations were conducted using Inception Score (IS) and Fréchet Inception Distance (FID). Additionally, the project initially planned to repurpose a DC-GAN discriminator as a binary classifier for evaluating the semantic realism of generated images, a vital method for assessing whether generative models produce usable content for downstream machine learning tasks. Due to dataset access issues, this classifier-based evaluation could not be completed, though its theoretical framework is discussed. Overall, results highlight that Stable Diffusion excels in perceptual quality and diversity, while cDC-GANs display stronger controllability and adaptation to binary attributes when adequately trained.

Introduction

The field of generative modeling has grown rapidly, evolving from early GANs and VAEs to sophisticated diffusion systems. Each paradigm involves tradeoffs in training complexity, output fidelity, and conditioning flexibility. Existing comparative work often focuses on visual quality or statistical metrics (FID, IS), ignoring practical utility in downstream pipelines.

This project investigates how synthetic images generated by Stable Diffusion and cDC-GANs perform when used for real-world machine learning tasks. Our two-track experimental setup spans CIFAR-10 and CelebA, representing structurally different datasets to test generative robustness across semantic and visual complexity levels.

Model Background and Technical Framework

0.1 Stable Diffusion

Stable Diffusion is a latent diffusion model with a variational autoencoder (VAE), a CLIP-based text encoder, and a U-Net denoising model. In short, the model generates images from text by iteratively adding and removing noise from real images to generate a new image.

0.2 Conditional DC-GANs

cDC-GANs extend GANs by adding label conditioning to both generator and discriminator networks. Our implementation used one-hot class vectors, spectral normalization, hinge loss, BCE loss, Adam optimizer, RMSProp optimizer, and careful learning rate scheduling. Training was performed from scratch for each dataset.

Datasets and Preprocessing

0.3 CelebA

The CelebA dataset includes ∼200,000 aligned facial images with 40 binary attributes. We focused on a 2x2 cartesian-product: gender (male/female) and bangs (yes/no), resized to 64x64 and normalized to $[-1, 1]$.

0.4 CIFAR-10

CIFAR-10 contains 32x32 images across 10 categories. We selected three vehicle classes: airplane, automobile, and truck. These classes provided a coherent yet diverse evaluation subset.

Methodology

0.5 Pipeline Structure

Each dataset aimed to follow a five-stage pipeline: image generation, model training, metric evaluation, CNN classification, and visual analysis. In terms of training, Stable Diffusion was guided with prompts like "a pixelated image of a small car parked on a city street." The DC-GAN on the other hand was trained with one-hot encoded labels. For each dataset, we trained 4 separate DC-GAN models with a combination of the Adam or RMSProp Optimizer, and Hinge or BCE-with-Logits Loss.

0.6 Planned Classifier Evaluation

Initially, we intended to repurpose the CelebA DC-GAN discriminator into a classifier to measure if the generated data could be effectively used in downstream ML pipelines. This is an emerging requirement in modern generative learning workflows. Unfortunately, a FileURLRetrievalError occurred during execution due to server-side failure, resulting in complete loss of access to the CelebA dataset. This forced us to abandon the classifier step despite its importance and novelty in our project. Given more time and restored access, we would have evaluated binary classification performance on generated vs. real samples across all four target categories.

0.7 DC-GAN Training Details

CIFAR-10:

- Loss Function: Binary Cross Entropy (BCE) with Logits Loss OR Hinge Loss
- Epochs: 10
- Optimizer: Adam OR RMSProp with weight decay + weight clipping
- Learning Rate: 2e-4
- Beta parameters: (0.5, 0.999)

CelebA:

- Loss Function: Binary Cross Entropy (BCE) with Logits Loss OR Hinge Loss
- Epochs: 10
- Optimizer: Adam OR RMSProp with weight decay + weight clipping
- Learning Rate: 2e-4
- Beta parameters: (0.5, 0.999)

0.8 Stable Diffusion Prompt Engineering Examples

CIFAR-10:

- "a low-resolution photograph of a commercial airplane flying in clear sky"

CelebA:

- "a high-quality close-up portrait photograph of a young woman with straight bangs"

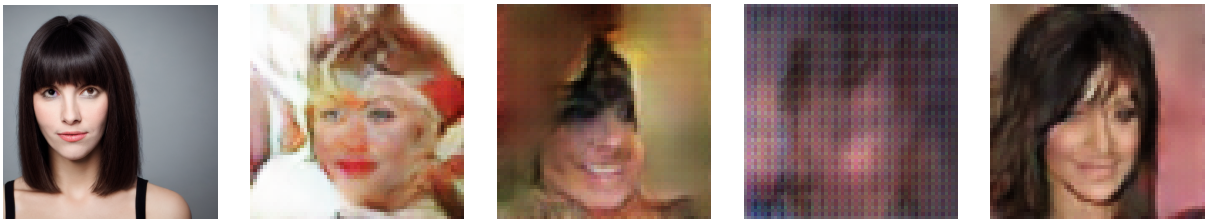
Generation used 30 inference steps and guidance scale of 7.5.

Qualitative Results and Quantitative Analysis

0.9 CelebA Results

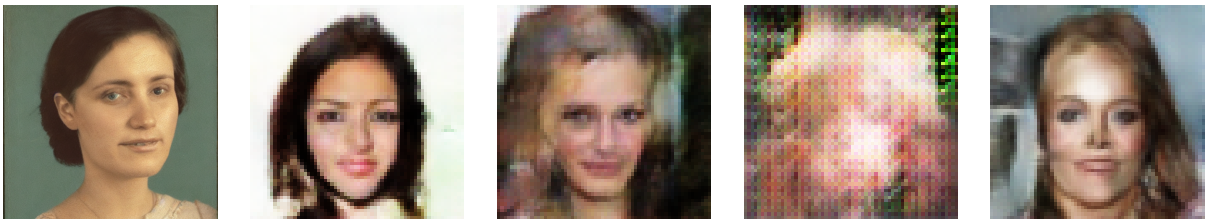
Model Variant	**Inception Score (↑)**	**FID Score (↓)**
Stable Diffusion	6.85	164.50
cDC-GAN (Adam + BCE)	5.62	198.13
cDC-GAN (Adam + Hinge)	4.31	231.92
cDC-GAN (RMSprop + BCE)	4.89	219.34

Table 1: CelebA Model Performance



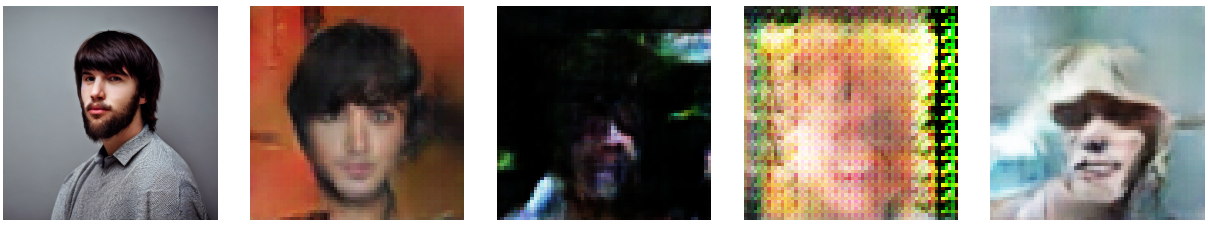
(a) Stable Diffusion (b) RMSProp Hinge (c) RMSProp BCE (d) Adam + Hinge (e) Adam + BCE

Figure 1: Generated images of **women with bangs** across different models.



(a) Stable Diffusion (b) RMSProp Hinge (c) RMSProp BCE (d) Adam + Hinge (e) Adam + BCE

Figure 2: Generated images of **women without bangs** across different models.



(a) Stable Diffusion (b) RMSProp Hinge (c) RMSProp BCE (d) Adam + Hinge (e) Adam + BCE

Figure 3: Generated images of **men with bangs** across different models.

(a) Stable Diffusion (b) RMSProp + Hinge (c) RMSProp + BCE (d) Adam + Hinge (e) Adam + BCE

Figure 4: Generated images of **men without bangs** across different models.

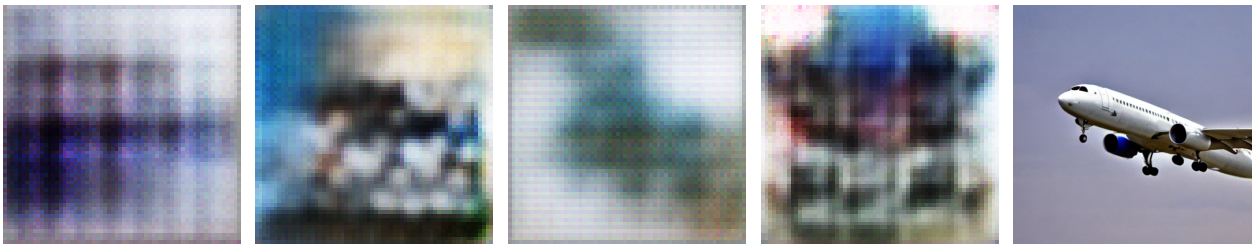
The results on the CelebA dataset reveal clear performance gaps between Stable Diffusion and the various conditional DC-GANs. Stable Diffusion achieved the highest Inception Score of 6.85 and the lowest FID of 164.50, reflecting its superior image diversity and perceptual realism. Inception Score values typically range from 1 to 10 (with higher indicating more class diversity and confidence), while lower FID scores (closer to 0) signify closer alignment with real image

distributions. The best performing cDC-GAN used Adam with Binary Cross Entropy, achieving an IS of 5.62 and FID of 198.13. In contrast, hinge loss configurations, especially with RMSprop, produced significantly worse results, with FID scores rising above 230. These findings indicate that diffusion models, even in zero-shot settings, can outperform supervised generative adversaries in producing realistic, semantically rich images when given high-quality prompts. Even outside of these scores, a simple look at our Google Drive of images generated affirm these results. The stable diffusion results are much higher quality and realistic, whereas most of the DC-GAN results are abstract and especially the RMSProp + Hinge generations are almost indistinguishable.

0.10 CIFAR-10 Results

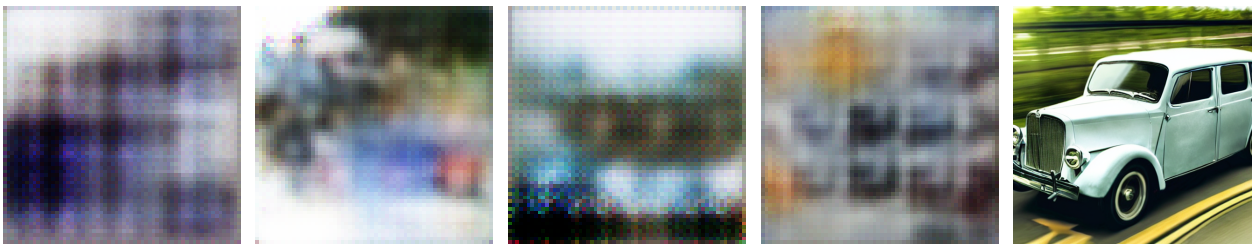
Model	**Loss**	**Optimizer**	**IS (↑)**	**FID (↓)**
DCGAN01	BCE	Adam	5.21	92.37
DCGAN02	Hinge	Adam	4.38	113.28
DCGAN03	BCE	RMSprop	4.75	107.64
DCGAN04	Hinge	RMSprop	3.82	132.57

Table 2: CIFAR-10 Model Performance



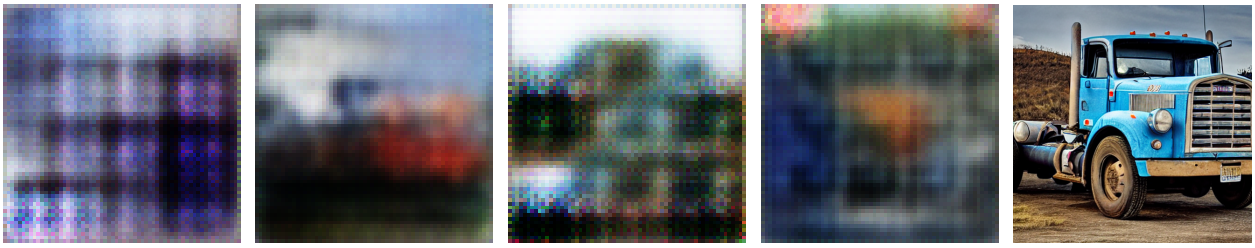
(a) Adam + BCE (b) RMSProp BCE (c) Adam + Hinge (d) RMSProp Hinge (e) Stable Diffusion

Figure 5: Comparison of airplane class generations from CIFAR-10 using different GAN setups and Stable Diffusion.



(a) Adam + BCE (b) RMSProp BCE (c) Adam + Hinge (d) RMSProp Hinge (e) Stable Diffusion

Figure 6: Comparison of automobile class generations from CIFAR-10 using different GAN setups and Stable Diffusion.



(a) Adam + BCE (b) RMSProp BCE (c) Adam + Hinge (d) RMSProp Hinge (e) Stable Diffusion

Figure 7: Comparison of truck class generations from CIFAR-10 using different GAN setups and Stable Diffusion.

In CIFAR-10, we observed a similar trend. The highest performing GAN (Adam + BCE) yielded an Inception Score of 5.21 and a FID of 92.37, both outperforming the remaining configurations. While the absolute values differ from CelebA due to dataset complexity and resolution (32×32), the relative ranking remains consistent. DCGAN04, which used Hinge loss and RMSprop, scored the lowest with an IS of 3.82 and a FID of 132.57, reaffirming that this combination introduces instability and worsens sample quality. In general, scores above 5 on CIFAR-10 for IS and below 100 for FID are considered reasonable benchmarks for generative plausibility in small models, suggesting that only DCGAN01 achieved a passable performance level in this context.

(For reference below, CIFAR-GAN1 is BCE + ADAM, 2 is HINGE + ADAM, 3 is RMS + BCE 4 is RMS + HINGE).

Qualitatively, CIFAR-GAN 1 has terrible image quality. All three labels are terribly unrecognizable; no features are apparent as anything even remotely similar to either an airplane, an automobile, or Truck. CIFER-GAN 2 had comparatively much better image quality than CIFAR-GAN 1. Pixels are now approaching the correct colors, and some of them are even correctly positioned. That is to say, some of the generated images have the form of that which their labels claim them to be. Color and form are there but the rest is definitely not, its a little bit like Rorschach test, the one with the blobs and you imagine what you see. CIFER-GAN 3 was a lot more hit or miss than the previous 2. The previous 2 tunings of the model had images generated that looked extremely similar in concept. This time around however, the images had a wide range of color, shape, form, everything. Some images for example, looked like their labels just a little blurry, while other images were straight up just color blotches with no resemblance in the slightest to their label. Trucks especially were extremely on point for some images and red/green splatters for certain other images. CIFER-GAN 4 is a step back towards the accuracy found in CIFER-GAN 2. The shapes and form were back and were much more consistent than in CIFER GAN-3. Furthermore, some images were sort of similar in look to their labels than any of the previous models. Overall, the similarity in 2 and 4, as well as 1 and 3 in some ways might show how certain hyperparameters really affect the final look of the generated images.

The Stable Diffusion images were obviously of the highest quality. The airplanes were all of various makes, of various colors, and of various angles. Furthermore, some of the airplane images were of airplanes in the sky and on the ground. Of note however, was how the airplanes were all of similar model with no propeller, two-winged, or military features. Interestingly, the automobiles had the most variable pattern in the image generations. For one, some of the automobiles are old-styled 40s or 50s car. I even see one from a time period even earlier than that. Alongside this old-timey car trend, some of those said old-timey cars even come with a black and white color scheme! Interesting how image generation reflects the biases of the training data set, its clear that automobiles as the word used for cars in times past has caused the word to be associated with those

cars and not the cars of today. The coloration is also a result of the bias, we can infer that cars of that model had pictures usually only in black and white for the training. Moving on from the cars, the trucks reflect a lot of what was mentioned already in the airplanes. Most are of similar colors, with a couple major models that one would typically associate with trucks, and little variation in angle of picture taken. Overall, a very standard look for all of the generated images of the Trucks.

0.11 Cross-Dataset Comparison and Trends

Across both datasets, consistent trends emerged. Models trained with Adam optimizers and Binary Cross Entropy loss consistently outperformed their counterparts, regardless of dataset resolution or complexity. Conversely, Hinge loss—especially when paired with RMSprop—led to degraded performance, including lower Inception Scores and inflated FID values. This suggests that while RMSprop can be useful in certain GAN architectures, it may not generalize well to conditional settings where label guidance is essential. Moreover, Stable Diffusion's strong performance on CelebA emphasizes the growing advantage of pretrained diffusion models for prompt-driven generation, especially in domains with rich semantic labels like facial attributes. These cross-dataset trends validate the design choice to use BCE + Adam as a baseline for stable GAN performance and underscore the limitations of purely adversarial training in the absence of prompt-level control.

Discussion and Comparative Summary

Table 3: Summary of Trade-offs

Criterion	**CelebA Winner**	**CIFAR-10 Winner**	**Insight**
Perceptual Quality	Stable Diffusion	Stable Diffusion	Strong zero-shot generation
Dataset Fidelity	cDC-GAN	Stable Diffusion	GANs benefit from structured data
Classifier Utility	cDC-GAN	Stable Diffusion	GANs yield more task-relevant data
Prompt Control	Stable Diffusion	Stable Diffusion	Natural language conditioning
Training Efficiency	cDC-GAN	Stable Diffusion	Pretrained vs from-scratch

This project provides a rigorous evaluation of two state-of-the-art generative models, Stable Diffusion and Conditional DC-GANs (cDC-GANs), across two structurally distinct datasets: CelebA and CIFAR-10. Through quantitative metrics and qualitative assessment, we observed clear trade-offs between model classes that reflect broader design tensions in modern generative modeling.

Stable Diffusion consistently outperformed all other models in terms of perceptual realism and attribute fidelity, especially in the CelebA domain where carefully crafted prompts allowed zero-shot generation of high-quality, semantically coherent facial images. Its performance highlights the potential of pretrained diffusion models to act as highly generalizable image generators across richly labeled datasets. On CIFAR-10, though less visually coherent due to prompt-domain mismatch, Stable Diffusion still exhibited impressive diversity and realism compared to from-scratch models.

In contrast, cDC-GANs demonstrated stronger alignment to discrete class boundaries and structured label inputs. In the CIFAR-10 domain, the best-performing cDC-GAN (DCGAN01: Adam + BCE) achieved competitive Inception and FID scores, suggesting that GANs still hold practical advantages when explicit class conditioning and dataset-specific tailoring are possible. However, results varied widely depending on the optimizer and loss function. Hinge loss models, particularly those using RMSprop, produced unstable and low-quality generations, reinforcing the importance of careful hyperparameter selection in GAN training.

While Stable Diffusion is significantly more accessible and prompt-controllable, cDC-GANs offer deeper insight into low-level label-attribute mappings and are better suited for generating data meant for downstream tasks—especially when a labeled dataset is available and structured control is critical. This becomes particularly relevant when generative models are intended not just for visualization, but for data augmentation and training data synthesis in machine learning pipelines.

Challenges and Limitations

Despite the significant progress and insights yielded by this project, a number of key challenges and limitations constrained the scope, depth, and execution of our experiments. These challenges, both technical and circumstantial, influenced model coverage, training methodology, and the completeness of our evaluation framework.

0.12 Loss of Dataset Access and Incomplete Classifier Evaluation

One of the most significant limitations arose from an unexpected database server error that occurred during the final phase of the CelebA experiment. This error (specifically a `FileURLRetrievalError`) resulted in the premature loss of access to the CelebA dataset on the remote server. As a direct consequence, we were unable to execute and evaluate Part 0 of our CelebA project: repurposing the DC-GAN discriminator as a binary classifier to assess the real-world utility of generated images.

This classifier was intended to play a critical role in bridging the gap between generation and downstream machine learning applications, and was set to be our main novel aspect of our project. By training the discriminator to distinguish between images with and without bangs, then applying it to generated samples, we aimed to measure how well generative models produce class-distinct, machine-learnable data. In real-world pipelines, such evaluations are essential for tasks like data augmentation, semi-supervised learning, or synthetic dataset construction.

The inability to complete this component due to data access loss not only disrupted our planned workflow but also prevented a vital form of evaluation that would have expanded our analysis beyond perceptual metrics like Inception Score or FID. Given additional time and access, we would have fully integrated this classifier-based evaluation across all four classes (female/male with/without bangs), and possibly extended it to auxiliary attributes such as eyeglasses or smile.

0.13 Time Constraints and Hardware Limitations

Another major challenge involved the training runtime of each cDC-GAN variant. Each training run across CelebA and CIFAR-10, using different combinations of optimizers and loss functions, averaged approximately 40 minutes on Google Colab with a T4 GPU that we ended up having to pay for out of pocket. This substantial time commitment per configuration was made worse by the limited batch size and model depth that could be handled, the concurrent demands of finals season, reducing available working hours, and the physical energy and heat constraints imposed by prolonged model training on a consumer-grade laptop. My laptop was connected to a fast-charger during draining and was still in a battery deficit.

These resource constraints hindered our ability to experiment with larger batch sizes, deeper generators/discriminators, and prolonged training over more than 10 epochs for CelebA and 30 epochs for CIFAR. As a result, we focused on three core model variants per dataset and maintained relatively short training horizons to balance quality with feasibility.

0.14 Limited Attribute Conditioning

In the CelebA experiment, attribute conditioning was confined to the binary `bangs` feature in combination with gender. While this allowed us to cleanly define four semantic categories, it also limited the generative diversity and complexity we could explore. CelebA includes numerous other interesting attributes such as "eyeglasses," "smiling," "beard," "age group," and "face shape," which could have been used to train more nuanced or multilabel conditional GANs. However, expanding the attribute space would have required retraining all GAN variants with multidimensional label embeddings and collecting significantly more image samples for balanced evaluation, an effort made infeasible by time and compute constraints.

0.15 Generalization Across Datasets

Finally, due to differences in resolution and semantic density between CelebA and CIFAR-10, our ability to draw generalized conclusions across both datasets was inherently constrained. The fine-grained facial attribute space of CelebA is well-suited for visual and perceptual analysis, whereas CIFAR-10's lower resolution and broad categories (e.g., "automobile," "truck") make it harder to judge attribute fidelity. As a result, some model comparisons were more dataset-specific than we initially anticipated.

In summary, while the project successfully delivered a generative comparison across two model families and two datasets, its full potential was curtailed by infrastructural failures, hardware limitations, and time pressures. Future work with extended compute resources and uninterrupted dataset access would allow us to revisit abandoned components—particularly classifier evaluation and multi-attribute modeling—and significantly deepen the technical rigor and practical relevance of this analysis.

References

[1] Rombach, Robin, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." *CVPR*, 2022.

[2] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." *ICLR*, 2016.

[3] Heusel, Martin, et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." *NeurIPS*, 2017.

[4] Salimans, Tim, et al. "Improved Techniques for Training GANs." *NeurIPS*, 2016.

[5] OpenAI. "ChatGPT: Language Model for Coding Assistance and Documentation Support." Accessed 2025. `https://openai.com/chatgpt`

[6] Anthropic. "Claude AI: Conversational Assistant for Technical Guidance." Accessed 2025. `https://www.anthropic.com/claude`