GenAi Masters At NaScon 2025 Report

Team: We3bears
Members: Moawiz, Areeba, and Sarim

April 20, 2025

Abstract

This project presents a creative three-phase pipeline combining generative AI and artistic interpretation. First, abstract social and environmental prompts were visualized using the Flux Schnell fp8 diffusion model. Next, a custom CycleGAN architecture was trained to learn artistic styles from a small dataset. Finally, the GAN was applied to the diffusion-generated images for stylized reinterpretation. Evaluation using CLIP, BLIP, FID, and LPIPS metrics highlighted how diffusion and GAN models can complement each other to create semantically rich and visually compelling art.

1 Phase 1 – Visual Interpretation with Diffusion Models

In this phase, we explored the hidden meanings behind cryptic phrases addressing social, health, or environmental issues. Using the **Flux Schnell fp8** model via the community-powered platform artbot.site, we generated visually symbolic representations of these themes.

The phrase that was given to us was this: "A Mind Lost in Load Shedding"

Model Details

Parameter	Value
Model	Flux Schnell fp8
Sampler	k_euler_a
Aspect Ratio	Square (1024x1024)
Steps	4
Guidance	1
Clip Skip	1
Platform	artbot.site/create

Table 1: Diffusion model configuration used for generating Phase 1 visuals.

Generated Visual Interpretations

P: Close-up of a man's face leaning toward a flickering lantern, his face half-lit, lips silently counting seconds
N: watch, digital clock, noisy background

P: A boy looking into a lantern as if seeking answers inside it, his eyes reflecting the flame like distant suns
N: fire hazard, electricity, superhero vibe

P: An elderly man lighting a candle at a family photo, the room around him silent and dust-laden
N: polished interior, modern furniture

P: Extreme close-up of an elderly man's wrinkled face, eyes closed in prayer, flickering lantern casting soft shadows on his features
N: modern face lighting, flash photography

P: A man gazing at a moth circling a flame, eyes wide with wonder and fear
N: clean surroundings, sunlight, bug zapper

Figure 1: AI-generated images interpreting abstract phrases using the Flux Schnell fp8 diffusion model.

2 Phase 2 – Custom GAN Architecture

In this phase, we implemented a custom GAN architecture from scratch inspired by CycleGAN, tailored to learn a specific visual style from a small dataset. The model was trained using [PyTorch] with optimizations suited for low-data regimes. Below are two sample outputs from different domains of the cycle that complement each other and demonstrate the model's learning capabilities.

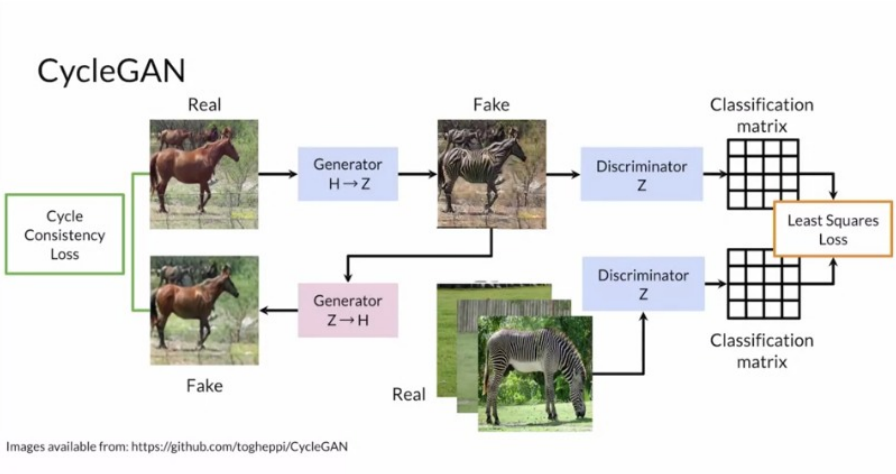


Figure 2: High-level CycleGAN architecture.

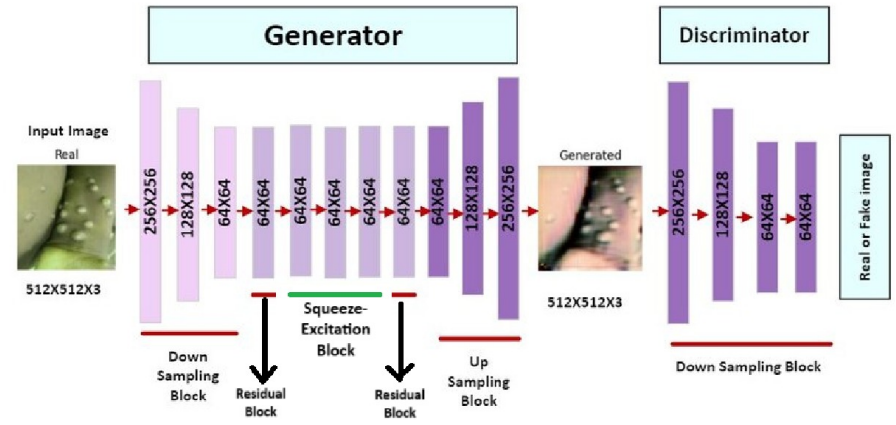


Figure 3: Custom Generator and Discriminator architecture of CycleGAN.

3 Phase 3 – Stylization: GAN + Diffusion Fusion

In the final stage of the pipeline, the images generated through the diffusion model in Phase 1 were stylized using the trained GAN model from Phase 2. This phase serves as the creative fusion of conceptual abstraction and learned visual aesthetics. The GAN, having learned a specific artistic style (e.g., cartoon, sketch, or painting), was now capable of applying this style to unseen inputs—in this case, the diffusion-based visual interpretations.

This transformation preserved the semantic core of the original phrases while enriching them with stylistic features. Below are the stylized outputs generated by the GAN for

three different concept-to-style applications:

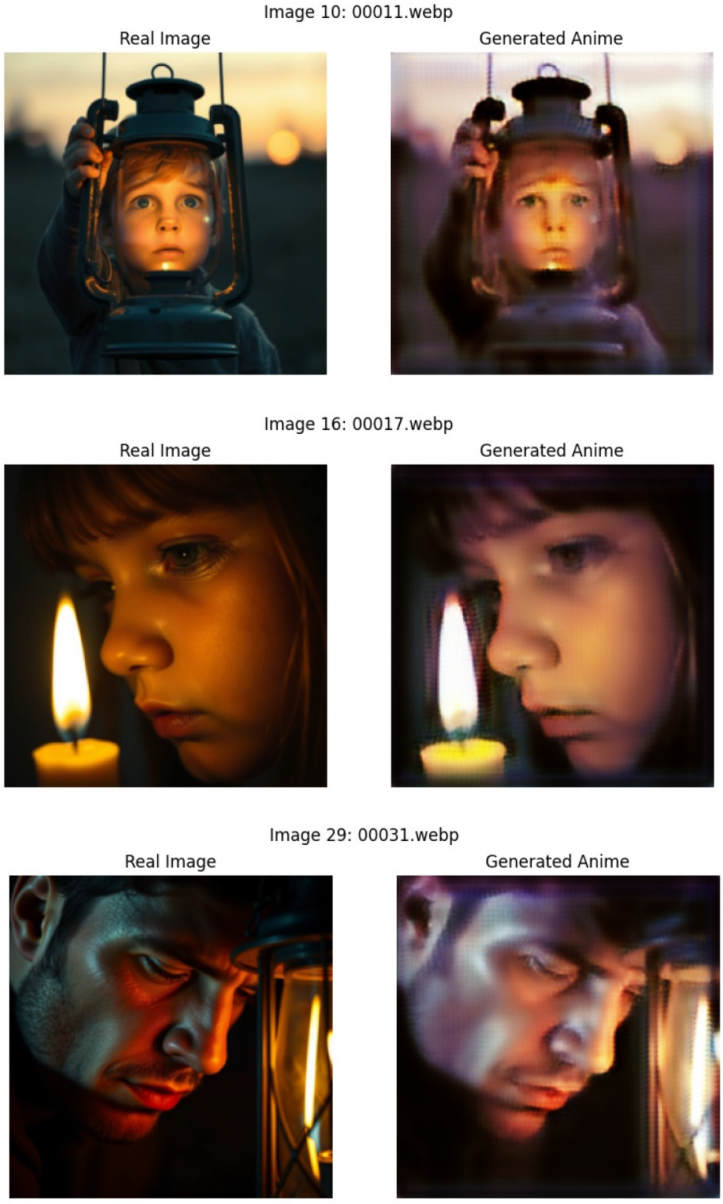
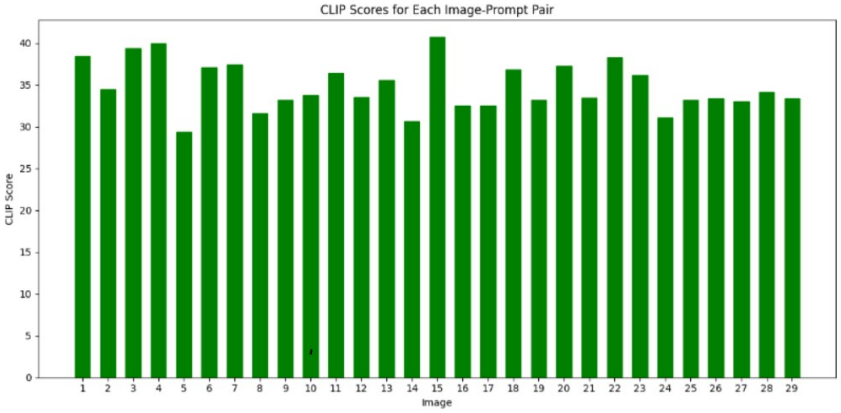


Figure 4: Stylized outputs from Phase 3: The GAN stylizes the diffusion-generated canvas images based on the learned visual theme, resulting in an expressive fusion of abstract interpretation and artistic transformation.

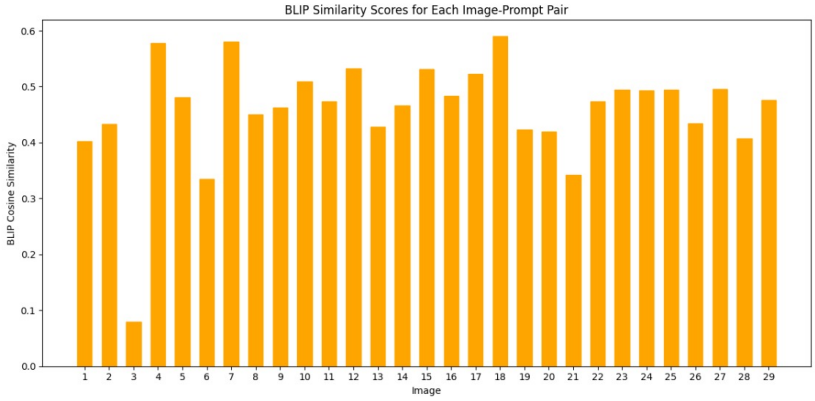
4 Evaluation

The performance of the generated images was evaluated using two key metrics:

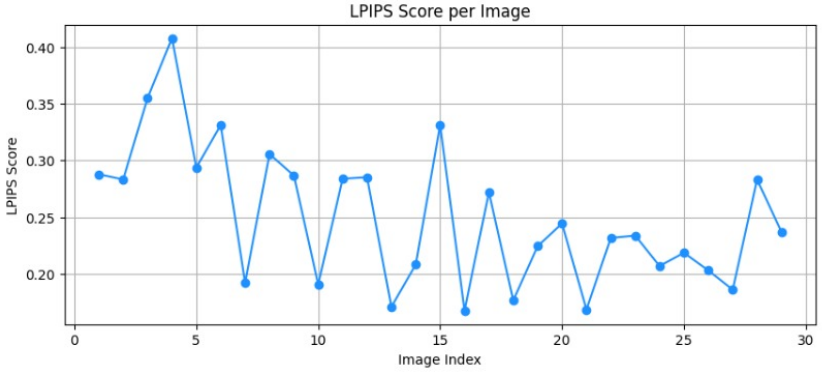
- **CLIP Score**: Measures alignment between the generated image and the input prompt.
- **BLIP Score**: Assesses image–text matching using the BLIP framework.



(a) CLIP Score Plot Across the Generated images



(b) BLIP Score Plot Across the Generated images



(c) Average FID Score: 269.5910, LPIPS Score: 0.2505

Figure 5: Comparison of CLIP and BLIP scores for text–image alignment.

CLIP BLIP Scores

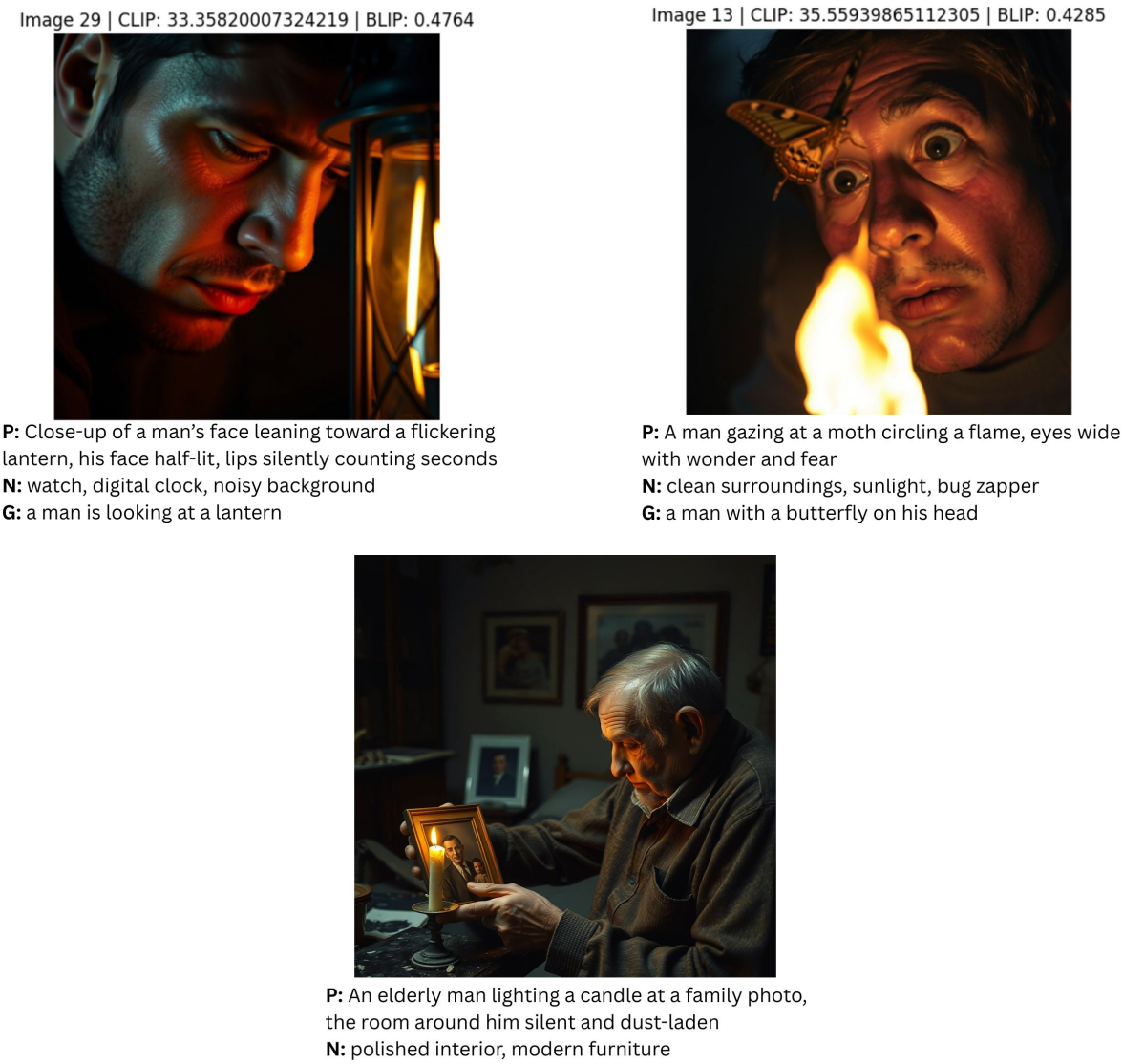


Figure 6: Visual interpretations including clip, blip score along with positive, negative and blip generated prompt.

FAQs

1. **Which Diffusion Model did you use and why? Did you use Quantization?**
   We used the **Flux Schnell fp8** model available on ArtBot, combined with the `k_euler_a` sampler. This model was selected for its high-quality outputs with minimal steps, making it well-suited for interpreting abstract phrases under limited computational constraints. **FP 8 Quantization** model variant was used to generate results faster.

2. **Explain your custom GAN architecture. Any innovations, optimizations, or mixed precision techniques?**
   We implemented a modified **CycleGAN** architecture using PyTorch. Our key

innovations included switching to Squeeze-Excitation blocks on Layer 2,3,4 of the Residual Network Blocks

This enabled our GAN to adapt better to the small style dataset and produce sharper, more consistent stylized images.

3. **Is it possible to train your architecture on multiple style datasets simultaneously? Why or why not?**

   While technically possible, training our current GAN on multiple style datasets simultaneously would lead to **style blending and mode collapse**, especially due to the limited size of each dataset. Without explicit style labels or conditional inputs, the model cannot distinguish or preserve distinct styles, making such training suboptimal. A conditional GAN (e.g., StyleGAN with style embeddings) would be better suited for this task.

4. **Is there a better approach for style transfer than GANs? Why?**
   Yes, **Diffusion Models** and **Neural Style Transfer (NST)** have shown to outperform GANs in certain aspects:

   - Diffusion models are more stable and produce more coherent textures and fine-grained styles.
   - Pre-trained diffusion-based stylization (e.g., ControlNet or StableStyle) can generalize across styles with minimal training.
   - NST techniques using feature maps from VGG can also transfer artistic styles with high fidelity.

   GANs are faster at inference but often suffer from instability during training and style leakage.

5. **Where did you observe limitations of GANs compared to diffusion models in this competition?**
   GANs struggled with:

   - Maintaining global structure across stylized images.
   - Learning complex styles with limited data (overfitting and artifacts).

   In contrast, the diffusion model in Phase 1 produced diverse and semantically rich visuals with just a few steps. This highlighted diffusion's strength in understanding prompts and preserving visual semantics better than our GAN.

Conclusion

This project demonstrated the powerful synergy between diffusion models and GAN-based style transfer for visual storytelling. By first interpreting abstract social or environmental phrases using a pre-trained diffusion model and then applying a custom-trained GAN to stylize those interpretations, we achieved visually compelling and semantically

rich outputs. Despite the limitations of GANs in capturing fine-grained details compared to diffusion models, the end-to-end pipeline showcased the potential of combining generative approaches for creative and meaningful visualizations.