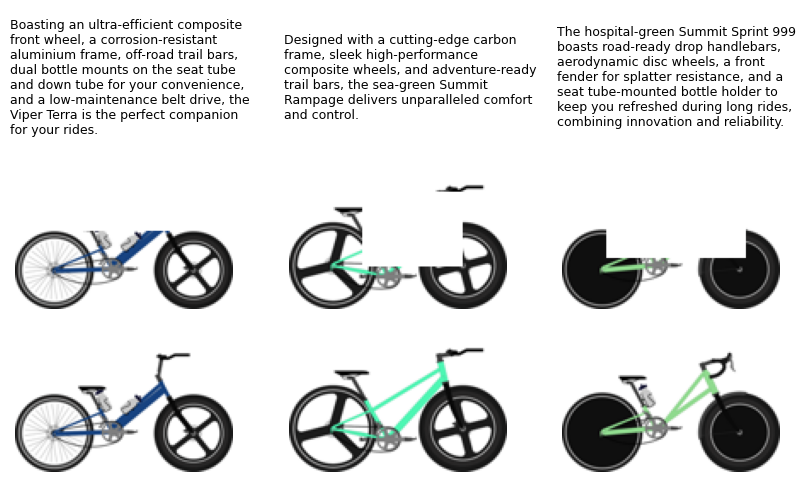
Code:

<https://colab.research.google.com/drive/1OxX4qKHPNC9aoY-yISDRxw_8u_JEEWQT?usp=sharing>

### **Bike Design Completion**

**Objective:** Automate bike design completion from partial images, descriptions, and parametric data.



**Introduction:**  
This project addresses the challenge of completing bike designs from partial inputs, marketing descriptions, and parametric data using generative AI. The goal is to automate the design completion process, enabling rapid visualization of creative concepts while maintaining coherence with the provided inputs. The dataset includes 10,000 training samples and 1,000 test samples comprising partial images, masks, descriptions, and ground truth targets.

Given-

**Partial Images**: Incomplete design sketches of bikes.

**Masks**: Highlighting the regions to be inpainted.

**Descriptions**: Marketing pitches describing the design intent.

**Parametric Data**: Numerical values representing design features.

**Targets**: Ground truth complete images for training.

**Test Partial Images** : The goal is to complete these Images.

**Test Description** : Test design description

**Test Parametric Data** : Test Parametric design data

| **Component** | **Description** | **Shape / Details** |
| --- | --- | --- |
| Partial Images | Incomplete design sketches of bikes | (10,000, 3, 128, 128) |
| Masks | Highlighting the regions to be inpainted | (10,000, 4) |
| Descriptions | Marketing pitches describing the design intent | List of 10,000 strings |
| Parametric Data | Numerical values representing design features | (10,000, n\_features) |
| Targets | Ground truth complete images for training | (10,000, 3, 128, 128) |
| Test Partial Images | Incomplete test images to be inpainted | (1,000, 3, 128, 128) |
| Test Descriptions | Test marketing descriptions | List of 1,000 strings |
| Test Parametric Data | Numerical values for test design features | (1,000, n\_features) |

Table 1: Dataset Overview

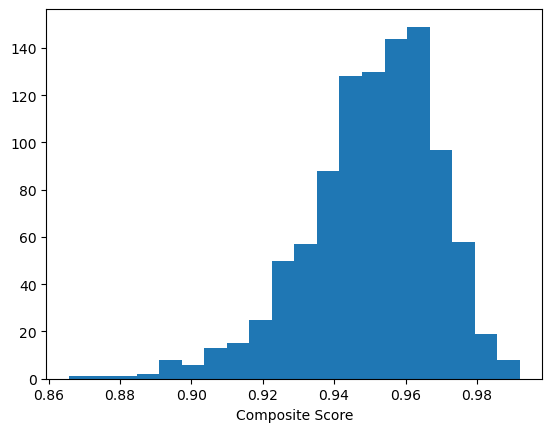
#### **Methodology**

1. **Models Explored:**
   * VAE, BikeFusion, and UNet.
   * Guidance mechanisms like CLIP embeddings and torch loss functions.
2. **Augmentation:** Applied transformations like rotations, skewing, and noise.
3. **Experimentation:** Fine-tuning and multimodal integration improved semantic alignment.

## **Approaches (ingonre max text in this section.. Try keeping graphs and image.. If not… leave this section)**

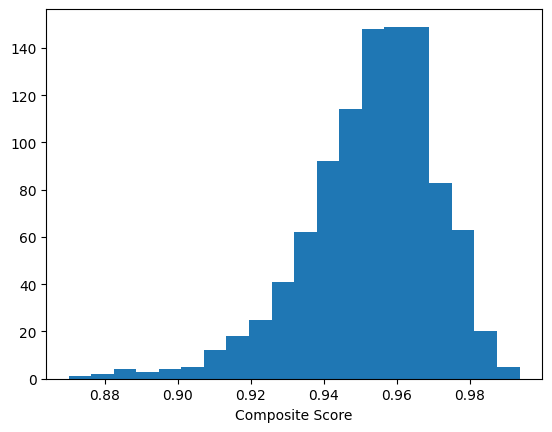
### 2.1. VAE (0.95063)

Variational Autoencoders (VAEs) are ideal for tasks that require **generative modeling with latent-space representation** and probabilistic interpretations. The model captures the overall distribution of the dataset but lacks targeted fine-tuning for task-specific guidance. However, it was a pretty high benchmark to even start with.



### 2.2. VAE + Data Augmentation (3X of given data) (0.95291)

Data was tripped using apply\_random\_mask function in Utils.py. The approach was to improve generalization and robustness by increasing training data variability. Augmentation exposed the model to diverse input scenarios, helping it generalize better. The next point here can be to apply even more types of masking like shape and sizes. Different rotations can also be used.

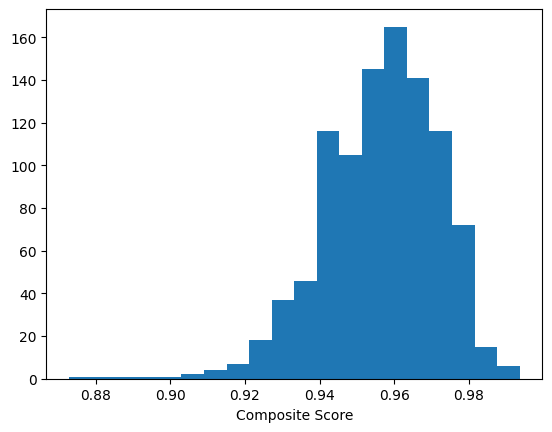


### 2.3. VAE + Data Augmentation with transformation (3X of given data) (0.956444)

This method applies diverse augmentations, including rotations, skewing, and Gaussian noise, to create a richer dataset with 3X the original data size. It builds upon the earlier data augmentation strategy by introducing more diverse transformations.

Achieved a composite score of **0.956444**, higher than the original VAE (0.95063) and the basic augmentation approach (0.95291).

The diversity in transformations helped expose the VAE to a wider range of input scenarios, enabling it to generalize better to unseen data. Some transformations, like extreme rotations or skewing, may distort the input beyond real-world plausibility, potentially introducing noise into the training.



| **Transformation** | **Description** | **Expected Impact** | **Observed Results** |
| --- | --- | --- | --- |
| Rotation | Random rotation within ±30 degrees | Exposes model to different orientations | Improved generalization |
| Skewing | Random affine transformations | Simulates diverse geometric distortions | Balanced robustness |
| Gaussian Noise | Random noise addition to mimic real-world variability | Increases noise tolerance | Enhanced reconstruction accuracy |

Table 2: Augmentation Strategies (2.3)

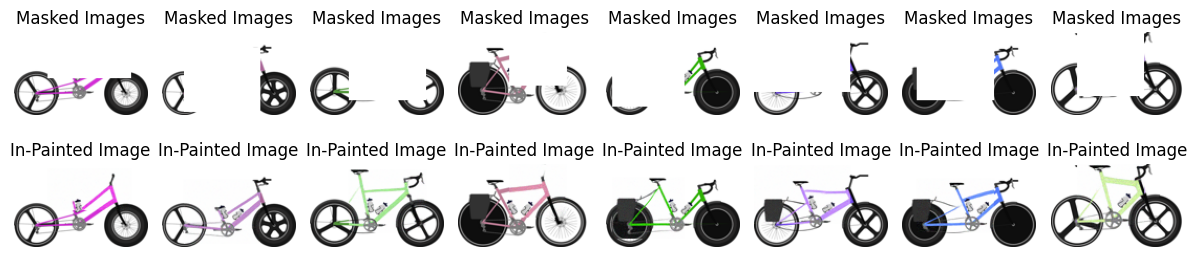
### 2.4. BikeFusion + Guidance 0.5 strength (torch.square) (0.95465)

BikeFusion model uses a guidance mechanism at 0.5 strength, with torch.square for computing differences to combine semantic guidance with image reconstruction for better predictions. The balance of guidance strength and torch.square loss helped align the generated images with descriptions and targets.

### 2.5. BikeFusion + Guidance 0.1 strength (torch.square) (0.95811)

guidance\_function=sample\_guidance

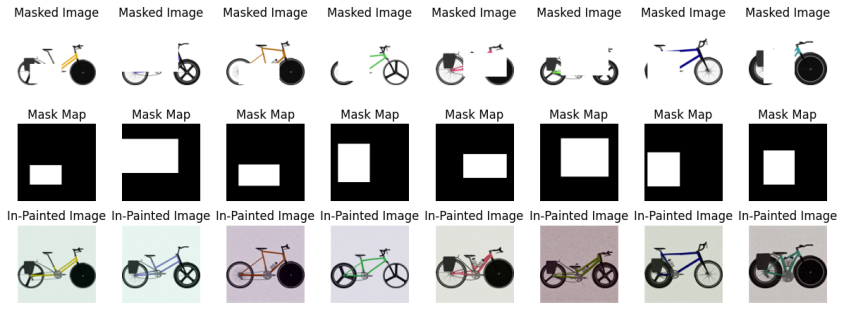
BikeFusion model with reduced guidance strength (0.1) and torch.square loss to test the impact of lower guidance strength on task performance. Lower guidance strength avoided overfitting to descriptions, enabling smoother blending of semantic and image features.

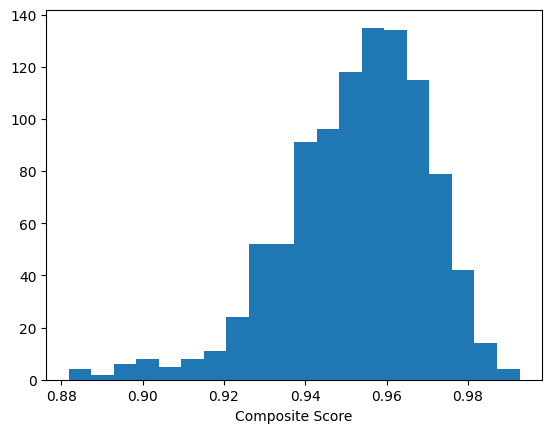


### 2.6. BikeFusion + Task-Specific Guidance (torch.nn.functional.mse\_loss) (0.919885)

guidance\_function=task\_specific\_guidance

BikeFusion model using task-specific guidance with torch.nn.functional.mse\_loss to incorporate direct loss computation aligned with the task's reconstruction goals. I felt this can be one of the best ways as Mean Squared Error (MSE) loss calculates the average squared difference between predicted and target values. It is a natural choice for tasks where accurate reconstruction is required, such as image generation, inpainting, and regression. MSE loss must have failed due to its inability to incorporate semantic understanding and perceptual alignment, which are crucial for our task.





### 2.7. BikeFusion + CLIP Guidance ( 0.955051)

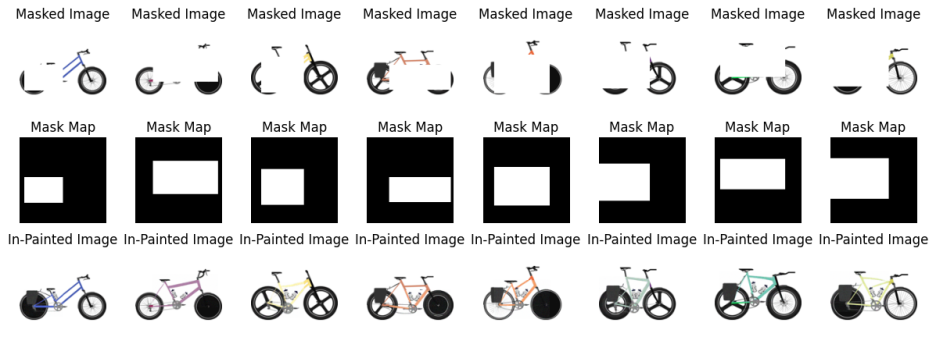
guidance\_function=text\_guidance

Combined BikeFusion's diffusion model with CLIP embeddings for multi-modal guidance align inpainting with textual descriptions to integrate semantic context from textual descriptions into the image inpainting process for better alignment with design prompts.

The strong text-to-image alignment capabilities of CLIP effectively steered the diffusion model toward generating more contextually accurate images. CLIP guidance (2.6) performed well but didn't outperform BikeFusion's 0.1 guidance (2.4) due to its over-reliance on textual semantics and lack of optimal fine-tuning for the specific task. The results highlight that balancing semantic guidance strength is crucial, as BikeFusion's 0.1 guidance demonstrated superior performance by effectively blending textual and visual features without overemphasis.

Also, it can be the CLIP doesn’t understand the “domain” relationship as shown by cosine similarity between image and accurate text prompt is just 0.2757.

Theoretically, I feel if we fine tune clip on a dataset of 10,000 pairs, we might get better results.



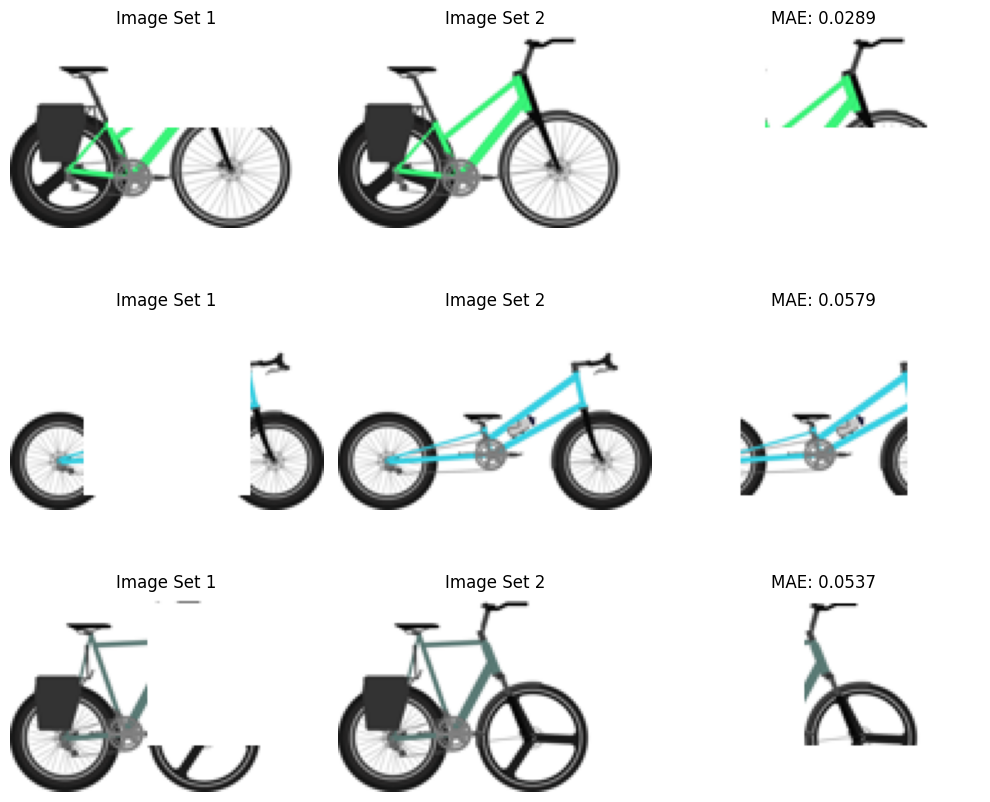
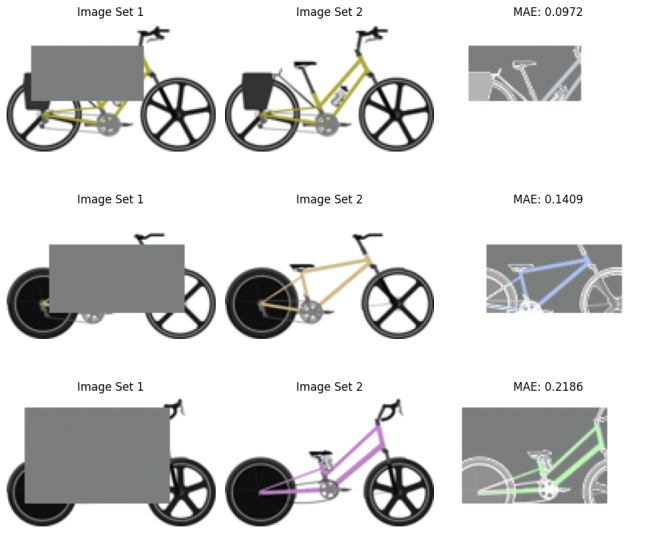
### 2.8. Unet

Convolutional neural network with encoder-decoder architecture, used for inpainting by reconstructing masked regions directly to explore a simpler, computationally efficient architecture compared to BikeFusion for image reconstruction tasks.

Performance was lower than BikeFusion with CLIP guidance due to the lack of multi-modal integration and the inability to leverage textual guidance.

UNet focused solely on pixel-level reconstruction without semantic understanding, limiting its ability to generate contextually relevant inpaintings.

### 2.9. GloVe + VAE



The GloVe + VAE approach combined semantic understanding from GloVe embeddings with the generative modeling power of VAEs to reconstruct images based on textual descriptions. While it effectively captured some contextual relationships, the alignment between text embeddings and image features was inconsistent, especially for complex masks or ambiguous descriptions.

This inconsistency resulted in higher reconstruction errors (MAE) for heavily masked images. Future improvements could involve using richer embeddings like CLIP(which I used.. Not so successful) or attention mechanisms to better integrate text and image data.

### 2.10. Stable Diffusion

The idea was to combine image and text data to fine-tune Stable Diffusion, aiming to make it domain-specific and better suited for generating accurate reconstructions. However, I found that direct fine-tuning of Stable Diffusion is not straightforward due to its architecture and pre-trained constraints. Additionally, the model's focus on creative generation rather than precise reconstructions made it less effective for our specific use case. This highlights the need for models more tailored to structured inpainting tasks.

## **Summary**

| **Approach ID** | **Methodology** | **Compos- te Score** | **Key Features/Remarks** |
| --- | --- | --- | --- |
| **2.1** | VAE (Given) | 0.95063 | Baseline method with probabilistic latent-space representation and generative modeling. |
| **2.2** | VAE + Data Augmentation (3X) | 0.95291 | Data tripled via random masking to improve generalization. |
| **2.3** | VAE + Data Augmentation with Transformations | 0.95644 | Applied diverse transformations like rotation, skewing, and Gaussian noise to improve robustness. |
| **2.4** | BikeFusion + Guidance 0.5 Strength (torch.square) | 0.95465 | Semantic guidance with a strength of 0.5 to balance description alignment and reconstruction. |
| **2.5** | BikeFusion + Guidance 0.1 Strength (torch.square) | **0.95811** | Lower guidance strength avoided overfitting and improved blending of textual and visual features. |
| **2.6** | BikeFusion + Task-Specific Guidance (torch.nn.functional.mse\_loss) | 0.919885 | MSE loss failed to incorporate semantic understanding and perceptual alignment. |
| **2.7** | BikeFusion + CLIP Guidance | 0.95505 | Multi-modal guidance combining CLIP embeddings with inpainting but lacked domain-specific fine-tuning. |
| **2.8** | UNet | - | Encoder-decoder architecture focused on pixel-level reconstruction, lacked multi-modal integration. |
| **2.9** | GloVe + VAE | - | Combined semantic understanding from GloVe embeddings with VAEs but struggled with alignment. |
| **2.10** | Stable Diffusion | - | Difficult to fine-tune and better suited for creative generation than structured inpainting tasks. |

## **Table 3: Approaches and Performance Comparison**

## **Discussion**

Nothing changed with this scheduler = CosineAnnealingLR(optimizer, T\_max=50, eta\_min=1e-5)

| **Methodology** | **Strengths** | **Weaknesses** | **Suggestions for Improvement** |
| --- | --- | --- | --- |
| VAE | Strong baseline for generative modeling | Lacks targeted task-specific guidance | Incorporate more targeted fine-tuning |
| VAE + Augmentation | Improved generalization through increased data variability | Limited diversity in augmentations | Introduce varied masking, rotations, and transformations |
| BikeFusion | Balanced semantic and image guidance | Over-reliance on guidance strength for specific performance | Fine-tune guidance strength dynamically |
| BikeFusion + CLIP Guidance | Integrated semantic context from text effectively | Domain knowledge in CLIP embeddings was suboptimal | Fine-tune CLIP on domain-specific dataset |
| UNet | Computationally efficient pixel-level reconstruction | Lacks semantic understanding, relying solely on pixel reconstruction | Add multi-modal capabilities to leverage textual descriptions |
| GloVe + VAE | Explored text-image integration through pre-trained embeddings | Inconsistent alignment between text embeddings and image features | Replace GloVe with richer embeddings like CLIP or use attention mechanisms |
| Stable Diffusion | Potential for creative text-to-image generation | Not optimized for structured inpainting tasks | Develop domain-specific variations of Stable Diffusion |

Table 4: Observations and Key Takeaways

| **Metric** | **Definition** | **Key Observation** |
| --- | --- | --- |
| Composite Score | Weighted combination of SSIM and MAE, adjusted for perceptual quality and alignment | Highest score achieved with BikeFusion + Guidance 0.1 Strength (0.95811). |
| SSIM | Measures structural similarity between reconstructed and target images | Augmentation and guidance improved SSIM. |
| MAE | Mean absolute error between predicted and ground truth pixels | Lower MAE observed with multi-modal approaches like CLIP Guidance. |

Table 5: Metrics Summary

1. **VAE as a Strong Baseline**: The VAE model performed surprisingly well, setting a solid benchmark with a score of 0.95063. However, it lacked any task-specific fine-tuning, which limited its ability to adapt to our specific needs.
2. **Power of Data Augmentation**: Adding more data through augmentation, such as random masks and transformations like rotation and noise, significantly boosted the model's performance to 0.95644. This showed how exposing the model to diverse scenarios made it better at generalizing.
3. **BikeFusion's Guidance Mechanism**: BikeFusion, combined with semantic guidance, struck a great balance between understanding text descriptions and reconstructing images. Reducing the guidance strength to 0.1 delivered the best results (0.95811), proving that subtle adjustments can have a big impact.
4. **CLIP Guidance Strengths and Weaknesses**: Using CLIP embeddings for guidance brought in semantic context from text, but its lack of domain-specific understanding held it back at 0.95505. Fine-tuning CLIP on a dataset of bike designs might unlock its full potential.
5. **Why MSE Loss Didn't Work**: Using task-specific MSE loss fell short (0.919885) because it couldn’t capture the semantic nuances of the task. This shows that reconstruction isn’t just about pixel accuracy—it needs to “understand” the design intent too.
6. **UNet's Simplicity and Limits**: UNet's simple encoder-decoder architecture worked well for straightforward inpainting but lacked the ability to integrate textual guidance. Without semantic context, it couldn’t generate designs as coherent as BikeFusion.
7. **GloVe Struggles with Alignment**: Combining GloVe embeddings with the VAE had potential but didn’t align text and images consistently, especially for complex or ambiguous descriptions. This highlights the need for richer embeddings or models like CLIP (fine-tuned for the task).
8. **Stable Diffusion’s Misfit**: Stable Diffusion, while great for creative tasks, didn’t suit our structured inpainting needs. It wasn’t easy to fine-tune for domain-specific tasks and struggled to generate accurate reconstructions.
9. **Composite Score as a Guide**: The composite score showed how well models balanced reconstruction accuracy (MAE) with perceptual quality (SSIM). Models that combined semantic guidance with inpainting consistently scored higher.
10. **Future Steps**: To push performance further, we could fine-tune multi-modal models like CLIP, explore dynamic guidance adjustments, and experiment with attention-based architectures.