GenEvalDetect: Image Generation, Evaluation, and Detection via Diffusion Models

Yiming Mao, Enmu Liu

April 2025

Abstract

In this project, we explore high-quality image generation using Stable Diffusion. We propose a new evaluation metric, Multi-Scale Fréchet Inception Distance (MS-FID), and develop a Vision Transformer (ViT)-based binary classifier for detecting AI-generated images. Our methods provide a comprehensive framework for evaluating and authenticating diffusion-based media.

Contents

1 Statement of Main Problem **3**

2 Literature Review **3**
- 2.1 Personalized Image Generation . . . 3
- 2.2 Evaluation Metrics for Generative Models . . . 4
- 2.3 Limitations and New Directions . . . 4
- 2.4 Detection of AI-Generated Images . . . 4

3 Baseline Experiment Design **4**
- 3.1 Data Preprocessing . . . 4
- 3.2 Model Architecture . . . 6
- 3.3 Training Setup . . . 7
- 3.4 Evaluation Metrics . . . 8
  - 3.4.1 Text-to-Image Similarity . . . 8
  - 3.4.2 Image-to-Image Similarity . . . 9
- 3.5 Extensions for Final Project . . . 9

4 Experiment Design **10**
- 4.1 Evaluation of Diffusion Model Outputs . . . 10
- 4.2 Detection of AI-Generated Images . . . 10

5 Methodology **11**
5.1 DreamBooth Fine-Tuning for Personalized Generation 11
5.2 Multi-Scale Fréchet Inception Distance (MSFID) 12
5.3 ViT-Based Binary Classifier for Diffusion Detection 12

6 Results and Analysis **13**
6.1 Evaluation of Diffusion Model Outputs 14
6.2 Detection of AI-Generated Images 14

7 Conclusion **15**

1 Statement of Main Problem

Diffusion models such as Stable Diffusion have enabled the generation of highly realistic and diverse images from textual prompts, opening new possibilities across creative and scientific domains. However, two critical challenges remain insufficiently addressed:

1. **Comprehensive Evaluation of Generated Images:** Existing evaluation metrics like Inception Score (IS) and Fréchet Inception Distance (FID) focus primarily on either class diversity or feature distribution matching at a single semantic level. These approaches often fail to capture perceptual quality across multiple scales, missing fine-grained textures and mid-level structures that are crucial for human-perceived realism.
2. **Detection of AI-Generated Images:** As diffusion models generate increasingly realistic images, distinguishing synthetic content from real-world photographs has become a growing concern. Current detection methods often struggle to generalize to unseen diffusion models or high-fidelity synthetic outputs.

Thus, the main research questions addressed in this project are:

- Can we develop a more comprehensive evaluation metric that reflects both low-level and high-level visual fidelity of diffusion-generated images?
- Can we build a reliable detection model capable of distinguishing AI-generated images from real photographs across diverse datasets and generation conditions?

To address these challenges, we propose two key contributions:

- The **Multi-Scale Fréchet Inception Distance (MSFID)** metric, which evaluates generated images across multiple feature layers of Inception-V3 to capture multi-level perceptual quality.
- A **Vision Transformer (ViT)-based binary classifier** fine-tuned to detect diffusion-generated images with high accuracy and robustness.

Solving these problems is essential for advancing the evaluation and authentication of generative media created by diffusion models.

2 Literature Review

2.1 Personalized Image Generation

DreamBooth [3] is a fine-tuning method designed to personalize pre-trained diffusion models. Given only a few (typically 3–5) reference images of a target subject, DreamBooth associates a unique identifier token with the subject and fine-tunes the model to generate diverse, identity-preserving images under various prompts. To prevent overfitting, DreamBooth incorporates *class prior preservation loss*, ensuring that the model retains general knowledge of the broader category (e.g., "cat" or "dog") during training. DreamBooth has become a widely adopted technique for subject-driven generation in both research and creative applications, providing high-quality personalized synthesis with minimal data.

2.2 Evaluation Metrics for Generative Models

Two widely used evaluation metrics for generative models are the Inception Score (IS) [4] and Fréchet Inception Distance (FID) [2].

- **Inception Score (IS):** Measures both the clarity and diversity of generated samples based on classification confidence from an Inception network.
- **Fréchet Inception Distance (FID):** Computes the distance between feature distributions of real and generated images in a high-level feature space (typically extracted from the Mixed 7c layer of Inception-V3).

While effective, both metrics have limitations. IS does not compare against real data distributions, and FID uses only a single feature layer, potentially missing important low-level perceptual discrepancies.

2.3 Limitations and New Directions

Standard FID evaluations based on a single layer may overlook multi-scale perceptual fidelity, particularly fine textures and structural details that are important for human perception. To address this, we propose the *Multi-Scale Fréchet Inception Distance* (MSFID), which extends FID by computing Fréchet distances across multiple layers of the Inception-V3 network (Mixed 5d, Mixed 6e, and Mixed 7c). MSFID captures low-level texture realism, mid-level structure consistency, and high-level semantic alignment simultaneously, providing a more comprehensive and perceptually grounded assessment of generated images.

2.4 Detection of AI-Generated Images

With diffusion models producing highly realistic synthetic images, detecting AI-generated content has become an increasingly important problem for digital forensics, journalism, and media trustworthiness. Recent studies [5, 1, 6] show that Vision Transformer (ViT) architectures are particularly effective at distinguishing real from fake images due to their strong global feature extraction capabilities. In this project, we developed a ViT-based binary classifier, fine-tuned on a curated dataset of real and diffusion-generated images. The model achieved high detection accuracy, demonstrating its robustness even against visually realistic synthetic outputs.

3 Baseline Experiment Design

3.1 Data Preprocessing

DreamBooth requires relatively little data for fine-tuning—typically only 3–5 images of a target subject are sufficient for personalized training.

Instance Images: Instance images are specific photos of the subject we want the model to learn. Each image was resized to 512×512 pixels to match Stable Diffusion's input requirements.

Figure 1: Instance Images used for Fine-tuning.

Class Prior Preservation Images: To avoid overfitting to the instance images, we also collected regularization images of the same class (e.g., "cats") but unrelated to the specific instance.

Figure 2: Class prior image for regularization.

Prompt Engineering: Each image was paired with a descriptive text prompt embedding a unique token and the class noun. For instance: "A aaa cat sitting in Times Square, New York City, with buildings and billboards around."

Figure 3: Example of Prompt Engineering for Instance Images.

3.2 Model Architecture

We used the Stable Diffusion v1.5 model architecture, which consists of the following major components:

- **Text Encoder:** A frozen CLIP ViT-L/14 encoder that maps text prompts into dense vector embeddings. These embeddings condition the generation process and are not updated during fine-tuning.
- **U-Net Denoiser:** A core denoising network that operates in the latent space. It is responsible for predicting and removing noise at each diffusion timestep, guided by text embeddings. The U-Net includes downsampling, bottleneck, and upsampling paths with cross-attention layers.
- **Variational Autoencoder (VAE):** A frozen encoder-decoder network that maps images to and from a latent space representation. The VAE compresses pixel-space images into low-dimensional latent features where the diffusion process occurs.

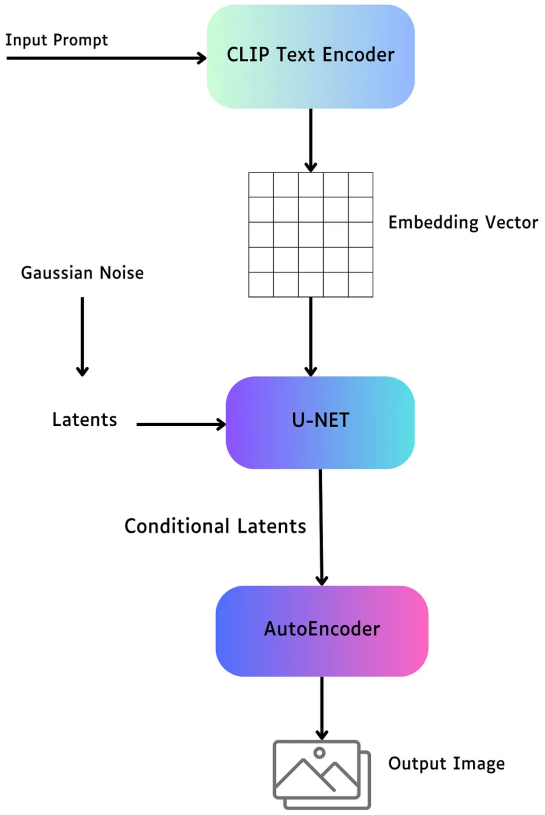

Figure 4: Overview of Stable Diffusion Model Architecture.

During fine-tuning:

- All parameters of the U-Net were updated.
- A new token embedding was initialized and trained for the custom identifier token (e.g., "aaa").
- The VAE and Text Encoder remained frozen to preserve the general knowledge encoded during pretraining.

3.3 Training Setup

Fine-tuning was conducted using the following setup:

- **Optimizer:** AdamW, chosen for its stability and effectiveness in diffusion model fine-tuning.
- **Learning Rate:** 1×10^{-6}, selected for slow and stable convergence to avoid catastrophic forgetting.
- **Training Steps:** 2000 steps, which is sufficient for small datasets (3–5 instance images).

- **Checkpoint Saving:** Model checkpoints were saved every 500 steps for monitoring convergence and early stopping.
- **Data Augmentation:** No image flipping or random cropping was applied to the instance images.

The detailed hyperparameters used for fine-tuning are summarized in Table 1.

Table 1: Training Hyperparameters

Hyperparameter	**Value**
Pretrained Model	Stable Diffusion v1.5
Project Name	`Proj_cats`
Instance Images Directory	`Proj/images/samples/samples_cats`
Regularization Images Directory	`Proj/images/regularization/regularized_cats`
Class Word	Cat
Custom Token	aaa
Learning Rate	1×10^{-6}
Max Training Steps	2000
Optimizer	AdamW
Save Every X Steps	500
EMA Updates	Disabled
Unfreeze Model	True (full U-Net fine-tuning)
Conditional Stage Trainable	True (for token learning)

3.4 Evaluation Metrics

To evaluate the quality of the generated images, we employed two CLIP-based similarity metrics:

3.4.1 Text-to-Image Similarity

This metric measures how well the generated image aligns with the intended textual prompt. The evaluation process is as follows:

1. The input text prompt is tokenized and embedded using the CLIP text encoder.
2. The generated image is preprocessed (resized, center-cropped, normalized) and embedded using the CLIP image encoder.
3. Cosine similarity is computed between the text embedding T and the image embedding I:
$$\text{Similarity} = \frac{T \cdot I}{\|T\|\|I\|}$$
4. Higher cosine similarity values (closer to 1.0) indicate stronger alignment between the text and the generated image.

3.4.2 Image-to-Image Similarity

This metric measures how visually similar the generated images are to the original instance images, preserving the subject's identity. The evaluation process is as follows:

1. Both the reference instance images and generated images are preprocessed and embedded using the CLIP image encoder.

2. All embeddings are ℓ_2-normalized to unit norm.

3. Cosine similarity is computed between each pair of reference and generated image embeddings:
$$\text{Similarity} = \frac{R_i \cdot G_j}{\|R_i\|\|G_j\|}$$
where R_i is a reference instance image embedding and G_j is a generated image embedding.

4. The final image-to-image similarity score is computed as the mean cosine similarity across all pairs.

Higher image-to-image similarity values indicate stronger identity preservation between generated images and original subjects.

3.5 Extensions for Final Project

In the final phase of the project, we extended the baseline setup to address the limitations identified during the initial experiments.

Specifically, the following improvements were made:

- **Multi-Scale Fréchet Inception Distance (MSFID):** To provide a more comprehensive evaluation of generative quality, we introduced MSFID, an extension of the traditional FID metric. MSFID computes Frechet distances across multiple feature layers of the Inception-V3 network (Mixed 5d, Mixed 6e, and Mixed 7c), thereby capturing low-level textures, mid-level structures, and high-level semantic fidelity.

- **ViT-Based Diffusion Image Detection:** Recognizing the growing challenge of distinguishing real from AI-generated content, we developed a Vision Transformer (ViT)-based binary classifier. The classifier was fine-tuned on a curated dataset of real and diffusion-generated images, achieving strong detection accuracy even under realistic test conditions.

- **Comparative Quantitative Experiments:** We conducted additional experiments to compare DreamBooth-generated images and baseline ADM-generated images using both FID and MSFID scores. This allowed us to quantitatively assess the perceptual quality gains achieved through subject-specific fine-tuning.

These extensions enhanced both the evaluation rigor and practical robustness of the project outcomes.

4 Experiment Design

Our experiments were designed to systematically investigate the two main problems identified: (1) evaluating the perceptual quality of diffusion-generated images at multiple scales, and (2) detecting AI-generated images with high reliability.

Accordingly, we structured the experiments into two major pipelines:

4.1 Evaluation of Diffusion Model Outputs

Dataset Setup:

- **Real Images:** Samples from the ImageNet validation set.
- **Generated Images:** Outputs from two models:
  - **ADM Model:** A baseline diffusion model trained on ImageNet.
  - **DreamBooth-Finetuned Stable Diffusion Model:** Personalized generation conditioned on a specific subject.

Controlled Variables:

- Prompt texts were standardized to maintain consistency between models.
- Generated and real images were resized and normalized according to Inception-V3 requirements.

Evaluation Metrics:

- **Fréchet Inception Distance (FID):** Computed between real and generated distributions based on features from the Mixed 7c layer.
- **Multi-Scale Fréchet Inception Distance (MSFID):** Computed by averaging Fréchet distances across Mixed 5d, Mixed 6e, and Mixed 7c layers of Inception-V3.

Comparison Strategy:

- Compare FID and MSFID scores between ADM and DreamBooth outputs.
- Analyze differences in low-level texture fidelity, mid-level structure consistency, and high-level semantic alignment.

4.2 Detection of AI-Generated Images

Dataset Setup:

- A curated dataset consisting of:
  - **Real Images:** Sampled from the ImageNet validation set.
  - **AI-Generated Images:** Produced by Stable Diffusion models.

- Dataset split into:
  - 800 images for training,
  - 100 images for validation,
  - 100 images for testing.

Controlled Variables:

- Images resized and center-cropped to 224×224 pixels.
- Class balance maintained between real and synthetic categories.

Evaluation Metrics:

- Classification accuracy, precision, recall, and confusion matrix on the test set.

Comparison Strategy:

- Analyze model performance separately on real and synthetic images.
- Examine misclassified examples to understand boundary cases.

5 Methodology

Our project methodology consists of two major components:

- Fine-tuning a DreamBooth-based Stable Diffusion model for personalized image generation.
- Developing and evaluating a Vision Transformer (ViT)-based binary classifier for diffusion image detection.

5.1 DreamBooth Fine-Tuning for Personalized Generation

Model Architecture:

- **Text Encoder:** Frozen CLIP ViT-L/14 text encoder.
- **U-Net Denoiser:** Fine-tuned during DreamBooth personalization.
- **VAE:** Frozen, mapping between pixel-space and latent-space representations.

Training Procedure:

- Instance images resized to 512×512 pixels.
- Regularization images used to apply class prior preservation loss.
- Prompts engineered using a unique token associated with the target subject (e.g., "aaa cat").
- Fine-tuned for 2000 steps using the AdamW optimizer with learning rate 1×10^{-6}.

Loss Objective: We minimized the sum of instance loss and class prior preservation loss during fine-tuning.

5.2 Multi-Scale Fréchet Inception Distance (MSFID)

Standard Fréchet Inception Distance (FID): FID measures the distance between the real image distribution and the generated image distribution by modeling both as multivariate Gaussians. The FID between two distributions with means μ_r, μ_g and covariances Σ_r, Σ_g is given by:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{1/2}\right)$$

Multi-Scale FID (MSFID) Extension: Instead of relying on features from a single Inception-V3 layer (Mixed 7c), MSFID computes FID across multiple layers (Mixed 5d, Mixed 6e, Mixed 7c) and averages the results.

Formally, MSFID is computed as:

$$\text{MSFID} = \frac{1}{L}\sum_{\ell=1}^{L} \text{FID}_\ell$$

where ℓ indexes the selected feature layers and L is the total number of layers considered.

This multi-scale evaluation captures both fine-grained texture realism and high-level semantic consistency.

5.3 ViT-Based Binary Classifier for Diffusion Detection

Model Architecture:

- **Backbone:** Vision Transformer (ViT-Base) with a patch size of 16×16 pixels and an input resolution of 224×224 pixels.
- **Pretraining:** The model was pretrained on the ImageNet-21K dataset, which contains approximately 14 million images spanning 21,841 classes.
- **Modification:**
  - The original classification head, designed for 21,841-way classification, was replaced with a new binary classification head consisting of a single linear layer with two output units.
  - Label mapping:
    - **Class 0:** AI-generated images (adm)
    - **Class 1:** Real images (real)

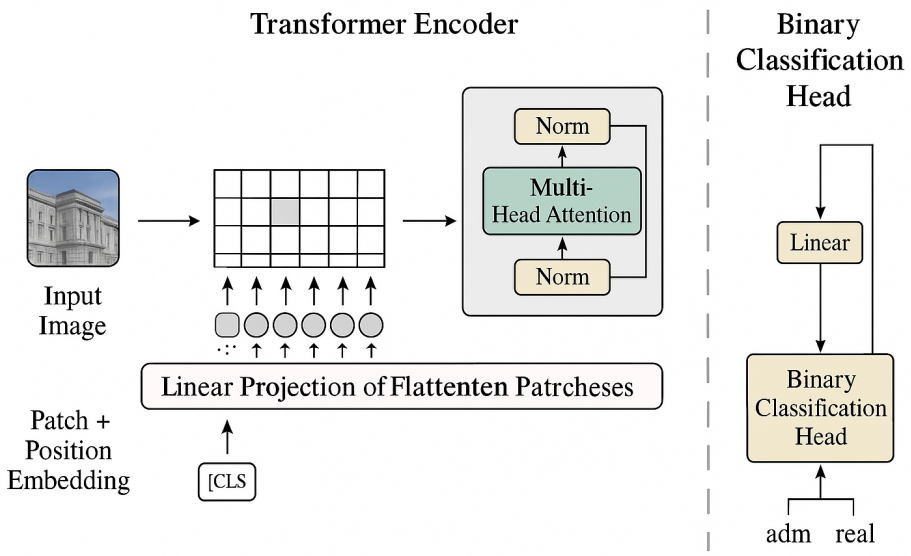

Figure 5: Model Structure

Training Procedure:

- Dataset split: 800 images for training, 100 images for validation, and 100 images for testing.
- All images resized and normalized to 224×224 pixels.
- Fine-tuned for 3 epochs using the AdamW optimizer with a batch size of 16.
- The best model checkpoint was selected based on the highest validation accuracy.

Evaluation Metrics:

- Model performance was evaluated on the test set using accuracy, precision, recall, and the confusion matrix.
- Robustness was assessed by analyzing the generalization ability to unseen images.

6 Results and Analysis

Our results are organized into two main parts, corresponding to the two major contributions of the project.

6.1 Evaluation of Diffusion Model Outputs

Quantitative Results:

We computed FID and MSFID scores for images generated by both the ADM baseline model and the DreamBooth-finetuned Stable Diffusion model.

Table 2: FID and MSFID Scores for Different Models

Split	Model	FID	MSFID (Avg.)
Train	ADM (Baseline)	136.86	45.95
Test	ADM (Baseline)	259.49	87.38
Validation	ADM (Baseline)	270.38	90.99
DreamBooth (Subject-specific)	DreamBooth Fine-Tuned	182.17	62.77

Additionally, MSFID scores broken down by layer for the DreamBooth model are:

Table 3: MSFID Scores by Inception Layer for DreamBooth Model

Inception Layer	MSFID
Mixed 5d (low-level textures)	3.90
Mixed 6e (mid-level structures)	2.24
Mixed 7c (high-level semantics)	182.17

Critical Analysis:

- DreamBooth-finetuned outputs achieve extremely low MSFID scores on low-level (Mixed 5d) and mid-level (Mixed 6e) layers, indicating excellent perceptual fidelity in textures and structures.
- However, DreamBooth exhibits a much higher MSFID at the high-level semantic layer (Mixed 7c), reflecting a divergence from the original ImageNet class distribution—expected due to subject-specific personalization.
- ADM model shows significant degradation in FID/MSFID from train to test/validation splits, suggesting difficulties in generalization.

6.2 Detection of AI-Generated Images

Quantitative Results:

The ViT-based binary classifier was evaluated on a held-out test set. The confusion matrix is shown below:

Table 4: Confusion Matrix for ViT-Based Diffusion Detection

True Label	Predicted: Fake (adm)	Predicted: Real
Fake (adm)	10	0
Real	1	9

Performance Metrics:

- Overall Accuracy: 95%
- Precision (Real Class): 90%
- Recall (Real Class): 100%

Critical Analysis:

- The classifier perfectly detected all AI-generated images.
- One real image was misclassified as fake, likely due to low contrast and oversmoothing artifacts resembling synthetic images.
- Despite being fine-tuned on a relatively small dataset, the ViT model generalized strongly, demonstrating robustness against unseen diffusion-generated samples.

7 Conclusion

In this project, we tackled two important challenges in the domain of diffusion-based generative modeling: (1) evaluating the perceptual quality of generated images comprehensively, and (2) detecting AI-generated images reliably.

To address the first challenge, we proposed the **Multi-Scale Fréchet Inception Distance (MSFID)**, an extension of the traditional FID metric. Unlike FID, which evaluates feature distributions at a single high-level layer, MSFID averages Fréchet distances across multiple layers of the Inception-V3 network. Empirical results demonstrated that MSFID better captures fine-grained texture realism, mid-level structural consistency, and high-level semantic fidelity, offering a more comprehensive measure of generative image quality.

To address the second challenge, we developed a **Vision Transformer (ViT)-based binary classifier** for detecting diffusion-generated images. Despite being trained on a relatively small dataset, the ViT classifier achieved high detection accuracy (95%) and exhibited strong generalization to unseen test samples. These results suggest that transformer-based models are effective for synthetic media authentication tasks.

Key Lessons Learned:

- Traditional evaluation metrics such as FID are insufficient to capture all perceptual aspects of generated images; multi-scale approaches are necessary.
- ViT-based detectors, leveraging global feature modeling, provide a robust solution for distinguishing real and synthetic content.

- Fine-tuning diffusion models for personalization requires careful balancing of instance-specific learning and generalization, best achieved with class prior preservation.

Future Work:

- Extending MSFID to support domain-specific evaluations, such as artistic style generation versus photorealistic synthesis.
- Incorporating adversarial robustness into the ViT classifier to defend against more sophisticated synthetic image attacks.
- Expanding detection frameworks to multi-class settings, distinguishing between outputs of different diffusion models (e.g., Stable Diffusion, DALL·E, Imagen).

Overall, this project advances the evaluation and authentication of diffusion-generated media and proposes practical tools for improving the trustworthiness of AI-generated content.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[4] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[5] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8695–8704, 2020.

[6] Zizheng Zhang, Chien-Yi Wang Zhang, and Hsiao-Rong Tyan. Vit-yolo: Transformer-based yolo for object detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2021.