Adobe Mid Prep Challenge
Image Classification and Artifact Identification

Report by - Team 73
Inter IIT Tech Meet 13.0

Figure 1: Is it Real or Fake?

Abstract

The widespread use of AI-generated imagery, fueled by sophisticated generative models e.g., BigGAN, GigaGAN, Stable Diffusion, etc., has made reliable and effective techniques for differentiating between artificially created and genuine pictures necessary, even for low-resolution or noisy images. [3] [9] [17]

As a solution for the first task (image classification), we propose a novel multi-head network designed to enhance classification accuracy and generalization in this scenario. Our approach leverages traditional CNN models as a backbone for efficient feature extraction, combined with a dual-head classification strategy which can predict the source (GAN or Diffusion) of the concerned fake image along with predicting real or fake. Our method also incorporates KANLinear layers [13] to generalize better on diverse datasets by leveraging its resilience against catastrophic forgetting, enhancing feature expressiveness without compromising efficiency. The model is refined for minimized size, decreased inference duration, and fewer parameters, facilitating scalable implementation. In addition, adversarial training that uses both authentic and synthetic images, along with defensive strategies, improves resistance to assaults and significantly reduces overfitting. Comprehensive trials demonstrate that our design achieves superior precision and resilient performance with no computing burden.

For the second phase of our pipeline, that is, artifact detection and reasoning, we employ a Vision Language Model (VLM) to furnish comprehensive justifications for the classification outcome. We utilize knowledge distillation to move knowledge from a larger, more intricate model to a considerably smaller, more efficient VLM. We meticulously refine the Llava v1.6 Vicuna 7B Vision [12] model utilizing a dataset of 2,338 photos tagged by GPT-4o, employing a human-in-the-loop methodology to guarantee superior precision and avoid hallucinations in annotations. This technique enables the smaller VLM to produce thorough and insightful explanations for each fake image without hallucinating or deviating from the provided possible artifacts, improving interpretability and reliability. By elucidating the rationale behind each choice, our methodology enhances comprehension of the model's behavior and cultivates more trust in its application for practical image forensics.

1 Introduction

In the recent years, with the rapid advancements in generative models, the generated images have become more and more realistic. Images ranging from landscapes (with broader details) to faces (with finer details) can now be generated with high similarity to real world images using various techniques. GANs (Generative Adversarial Networks), VAEs (Variational Autoencoders) [11], and Stable Diffusion models are some of the models which are being highly exploited and being developed on for the generation of images and improving the quality of generated images.

In parallel with the advancements in these generative techniques, the proliferation of fake images, including Deepfakes and other fabricated visuals, has led to challenges in image verification. Deepfakes, which utilize GANs to create hyper-realistic images or videos, have already been used for significant social and political risks, such as the spread of misinformation and manipulation. Furthermore, diffusion models, used for generating high-quality images, are also contributing to the growing concern over the authenticity of images shared online. These synthetic images often contains of subtle artifacts which are difficult for a human to detect but may get identified through different techniques using deep learning. Detecting these artifacts becomes important news verification, personal identification, law enforcement, etc., where the integrity of visual content is a need.

For example, GANs map random noise to a real image, creating a completely synthetic image which mimics the structure, texture, and lighting of real-world images. On the other hand, diffusion models work by gradually transforming random noise into a coherent image, which generally leads to more realistic and high-quality images when compared to GANs.

The rise of generative models thus has enabled the creation of fake images that are becoming increasingly difficult to distinguish from real ones and pose a problem to the society. The ability to detect AI generated images is becoming critical for maintaining trust in digital media day-by-day. Existing detection methods, have been successful for identifying images generated by older techniques like GANs, but they struggle to identify artifacts in newer, higher-quality models such as diffusion based generators.

These detection models often fails to give higher accuracies when trained on data generated from different kind of generative models and their variations. Therefore, there is a need for a robust solution that can accurately classify AI-generated images across a wide range of generative techniques.

In this first part of the problem statement, we need to develop a model capable of identifying AI-generated images and distinguishing them from real ones. The models are moreover required to perform classification of low quality images of size 32×32 pixels. The goal is not just classification but to achieve accuracy and efficiency, even for images originating from diverse generative models. The challenge here is in detecting even the subtle artifacts left behind by generative models, which may not be very visible in low-resolution images like the ones provided in the example dataset.

While accurate classification an image as real or fake is important, it is also important to provide interpretable explanations for these classifications by explaining the artifacts that led to the classification. Examples of artifactis in AI generated images includes unnatural blending, anatomical inconsistencies, color treatment abnormalities, resolution issues, unrealistic motion blur, etc. By creating a model that can give reasons, we can ensure greater transparency and trust in the detection.

To address the second task of the problem statement, our approach combines state-of-the-art methods in image classification and artifact detection. First, we develop a multi-head convolutional neural network (CNN) model that efficiently extracts features from low-resolution 32×32 images. The multi-head design enables the model to not only predict whether an image is real or fake but also to classify the source of the fake image (i.e., GAN or diffusion model). This dual classification capability enhances the model's ability to generalize across different types of fake images.

After the classification task, we incorporate a second phase focused on artifact detection and explanation generation. For this, we use a Vision Language Model (VLM) that can provide interpretative explanations for the artifacts detected in each image. This model is trained to detect distinguishing artifacts, and give an explanation for how these artifacts contributed to the classification, thus giving a trustable pipeline.

2 Related Works

The detection of AI-generated images has now become a very important area of research, especially since generative models such as GANs eg. *StyleGAN, BigGAN, GigaGAN, etc* , and diffusion models that can create highly photorealistic images have emerged. As the models continue to be perfected, early detection methods dependent on identifying irregularities in texture and biological anomalies will not be effective (Amoroso et al., 2023). Classifiers generalize very poorly on other generative models and perform abysmally on newer variants of them [15]. Furthermore, measures of local intrinsic dimensionality are also becoming ineffective because of the increased realism that images produced by generative models possess[14].

Reconstruction-based detection approaches, that rely on overfitting diffusion models, have proven to be effective but rapidly become intractable as the generators improve[4, 22]. Other strategies are enhancement of sub-artifacts or model boosting, which can only provide higher robustness, as high-quality deepfakes require greater computational powers.

Some methods attempt to enhance sub-artifacts by using techniques such as visibility matrices and model boosting robustness. Such methods, however, fail in front of higher quality deepfakes with additional computational costs. For example, hybrid architectures of neural networks, which combine InceptionResNetV2 with Conv2D layers, have proven efficiency in AI-generated image detection. However, such methods lack scalability in practical scenarios and fail to adapt to alternate datasets [18].

A critical gap in current research is the detection of low-resolution deepfakes, particularly those created at 32x32 resolution. While most studies focus on higher-resolution images (e.g., 512x512), little work has been done on detecting low-quality deepfakes that have been heavily compressed, a common occurrence on social media platforms that use varying compression factors and algorithms[5].

Also all these advances in current detection approaches face significant challenges regarding generalization, scalability in practical scenarios, and post-processing attacks. Most current approaches also lack interpretability, thereby limiting their use in high-stakes fields such as digital forensics. All of these challenges highlight the present necessity for more flexible, scalable, and transparent detection techniques that will be consistent with the fast-evolving generative models landscape.

3 Key Challenges

Some of the challenges we had to figure out are as follows:

- **Combining Data from Multiple Sources:** To make sure that the model is sufficiently robust, we had to make sure that we train it on data from various sources. For this, we had to combine data from GAN based models and diffusion based

Figure 2: high resolution (first row) versus low (32*32 pixel) resolution (second row) images

models. Despite our efforts, we were not able to incorporate data from certain important sources, such as GigaGAN, due to absence of publicly available dataset. The absence of such sources limited the model's exposure to diverse and improved generative techniques, potentially reducing its efficacy against advanced synthetic content.

- **Absence of Annotated Data for Fine-Tuning:** There was no reasoning based publicly available dataset. The lack of this data specifically tailored to the second task posed a significant challenge. Fine-tuning a pre-trained model or training a custom architecture often relies on labeled dataset with images and their annotated datasets. Without such data, we had to rely on Knowledge Distillation.[7]
- **Low-Resolution of Input Images:** The problem statement required us to work with images of size 32x32, which presented a significant challenge due to the limited amount of detail and texture that such small images can capture. This constraint not only made it harder to detect fine-grained artifacts in AI generated images but also limited the ability of the model to differentiate between subtle features of real and synthetic images. The smaller resolution restricted the visibility of patterns like texture irregularities, lighting inconsistencies, or fine details, which makes it difficult for the model to reason out the artifacts.
- **Adversarial Robustness:** Adversarial noise is often inserted in fake images to avoid detection. Thus, we had to ensure that the dataset we used for task 1 and task 2 both had examples of images which were adversarially attacked so that our model does not get fooled by such images during inference. [6]

4 Our Approach

4.1 Task-1 (Image Classification)

4.1.1 ***Dataset Preparation.*** We leveraged *GenImage*, a dataset specifically designed for million-scale AI generated image detection. The dataset comprises of over *one million pairs of forged and genuine images* from different generative models such as BigGAN, Stable Diffusion v1.4 and v1.5, Midjourney [2], Wukong, VQDM, etc. This dataset promises a good range of high quality realistic images.

For our experiments, we utilized a subset of images available in *GenImage* dataset, which had images generated by *BigGAN, Stable Diffusion, and MidJourney*. Additionally, we gathered more data by including *StyleGAN [10] generated images* from Kaggle and low-resolution images from the *CIFAKE* [1] dataset. This resulted in a highly curated set of over *0.5 million high-resolution images*, comprising both synthetic and real world image.

After curating the images, we filtered the images such that each image contained at max three distinguishable objects occupying approximately *10% of the total image area*. Fully automatic filtering was performed using *YOLOv10* [21], which efficiently selected images with a prominent presence of objects while minimizing the focus on large background areas in a single image. This was done to ensure that the training data has images with objects so that it can learn to identify features in objects rather learning on random backgrounds.

Downsampling. As we needed low resolution images for our problem statement, the high-resolution images were downscaled to *32x32 pixels* using bilinear interpolation, as per the dataset generation methodology described in the *CiFAKE* paper.
The formula used for bilinear interpolation is as follows:

$$\begin{aligned} f(x,y) =&(1-u)(1-v)f(x_1,y_1) + u(1-v)f(x_1+1,y_1) \\ &+ (1-u)vf(x_1,y_1+1) + uvf(x_1+1,y_1+1) \end{aligned} \quad (1)$$

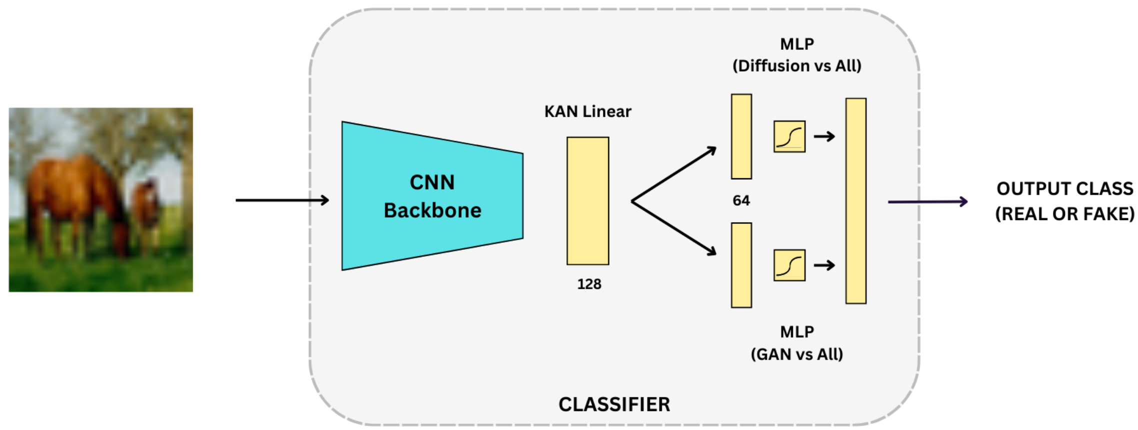

Figure 3: Classifier Architecture

Table 1: Classifier Dataset Overview

Dataset	No. of Fake Images	No. of Real Images
Stable Diffusion 1.5	64,938	36,257
BigGAN	120,000	161,996
StyleGAN	17,102	NaN
CiFAKE	50,000	50,000
Midjourney	59,527	67,785
TOTAL	**311,567**	**316,038**

Adversarial Training and Augmentations. Adversarial perturbations were applied to the images prior to downsampling using methods such as *FGSM* and *PGD* [20]. Additionally, several augmentation techniques were applied to further enhance the dataset. These augmentations improved the robustness of the models trained on the dataset, especially against adversarial attacks and random noise in natural images.

4.1.2 ***MultiHead Architecture-based Classification.*** We propose a MultiHead architecture to optimize the performance of image-based classification of real vs AI-generated images. Our primary architecture is *MultiHeadEfficientNet-B3*, which combines the efficiency of the *EfficientNet-B3* [19] backbone with a dual-head classification strategy. This approach was motivated by the limitations of a single CNN model to classify real and AI-generated images from diverse sources, such as *BigGAN, GigaGAN, StyleGAN, and Stable Diffusion*. Some key components of our architecture:

- *Backbone:* Our architecture primarily uses the *EfficientNet-B3* backbone as the core feature extractor because of its efficiency and performance. However, the framework is designed with modularity, allowing alternative models to be substituted in. This flexibility enables systematic experimentation with diverse architectures.
- *Flexible Feature Projection Layers:* After the image is passed through the convolutional backbone, the output feature maps are then flattened and passed through an embedding projection layer. Our architecture is designed to support both *KANLinear* layers and standard linear layers for feature projection. However, *KANLinear* layers demonstrate superior performance compared to standard linear layers. This advantage arises from the ability of KANs to learn more robust and generalizable features compared to traditional Neural Networks. KANs are less susceptible to *catastrophic forgetting*, a phenomenon where neural networks lose previously learned information when adapting to new tasks. Thus, KANLinear layers should be able to preserve knowledge in more complex scenarios like ours.
- *Separate Classification Heads:* In our architecture, the two heads are designed as independent experts, each specializing in a distinct classification task. One head is dedicated to distinguishing between real images and those generated using diffusion-based techniques, focusing on identifying the subtle artifacts and unique patterns characteristic of diffusion models. The other head specializes in differentiating real images from GAN-based fake images, leveraging its ability to detect the distinct features and inconsistencies commonly associated with GAN outputs. This dual-head specialization enables the model to effectively address both classification tasks with a focused and expert approach, ensuring robust performance across diverse synthetic image types.

4.1.3 ***Emphasis on Efficiency.*** We critically evaluated the trade-off between model performance and computational efficiency. Models with fewer parameters, lower CPU inference time, and smaller

Table 2: MultiHeadEfficientNetB3 Model Architecture.

Component	Description	Output Shape
Backbone	EfficientNet-B3 feature extractor (pretrained).	(B, C, H, W)
Flatten Layer	Flattens the backbone output to a 1D tensor.	$(B, flatten\_dim)$
KANLinear layer	KANLinear layer for shared feature embedding.	$(B, 128)$
Class 1 Branch		
fc1_class1	Fully connected layer ($128 \rightarrow 64$).	$(B, 64)$
dropout_class1	Dropout layer for regularization ($p = 0.2$).	$(B, 64)$
fc2_class1	Fully connected layer for logits ($64 \rightarrow 1$).	$(B, 1)$
drop_class1	Dropout layer for regularization ($p = 0.5$).	$(B, 1)$
Class 2 Branch		
fc1_class2	Fully connected layer ($128 \rightarrow 64$).	$(B, 64)$
dropout_class2	Dropout layer for regularization ($p = 0.2$).	$(B, 64)$
fc2_class2	Fully connected layer for logits ($64 \rightarrow 1$).	$(B, 1)$
drop_class2	Dropout layer for regularization ($p = 0.5$).	$(B, 1)$

Table 3: Classifier Model Efficiency Comparison

Model	Number of Parameters	Inference Time (ms)	Accuracy (%)
EfficientNet-B3	11.08 M	75	82.1
EfficientNet-B3 + KANLinear	12.68M	85	85
ResNet18	11.80M	95	78.7

model sizes were prioritized. These considerations align with the practical need for deploying efficient and scalable models in real-world applications, where inference speed and model size significantly affect user experience and operational efficiency.

Table 4: Accuracy comparison on our data with EfficientNet B3 backbone

Configuration	Parameters	Accuracy
With KANLinear	12.68 M	85%
Without KAN	11.08 M	83.1%

4.1.4 *Training and Evaluation Framework*.

For all documented architectures, we used a consistent shared training framework to ensure consistent evaluation. The key elements of this framework include:

- *Loss Function:* Binary Cross-Entropy (BCE) loss, applied individually on the results of both the heads.
- *Optimizer and Scheduler:* Adam optimizer with a `ReduceLROnPlateau` learning rate scheduler.
- *Mixed Precision Training:* To accelerate convergence and reduce memory overhead, we utilized PyTorch's mixed precision training.

4.1.5 *Model Prediction and Classification Logic*.

Our architecture outputs probabilities from its dual-head configuration. These probabilities are post-processed to determine the final classes of input images: *Real*, *Diffusion Generated Image*, and *GAN Generated Image*.

Table 5: Hyperparameter Settings and Training Details

Category	Details
Loss Function	Binary Cross-Entropy Loss (`torch.nn.BCEWithLogitsLoss()`)
Optimizer	Adam
Learning Rate (`lr`)	0.001
Weight Decay	1e-5
Learning Rate Scheduler	ReduceLROnPlateau
Mode	Min
Factor	0.1
Patience	3
Gradient Scaling	Mixed Precision (`torch.amp.GradScaler('cuda')`)
Early Stopping	
Patience	10 epochs
Best Validation Loss	Initialized to infinity
Training Settings	
Epochs	Set by `num_epochs`
Device	A5000 GPU (`torch.cuda`)
Classification Threshold	0.5

For final binary classification, both Diffusion and GAN-generated images are classified as *fake*. The intermediate classification (three classes) is based on a threshold-based classification logic that incorporates:

- Independent predictions from both heads.
- A decision layer to combine outputs from the heads into a single prediction. *The decision layer checks if any of the head classified the image as fake, if so, it classifies the image as fake and if both classifies the image as real then it is classified as a real image.*

4.1.6 ***Advantages of Our Approach***.

- *Flexibility:* Incorporates diverse backbones while maintaining efficiency.
- *Generalization:* Tested multiple configurations to reduce the risk of overfitting.
- *Efficiency:* Emphasis on reducing parameters and optimizing inference time.
- *Lower Resolution:* Our method works reasonably well on 32 x 32 resolution images (85 ± 2% accuracy on our compiled dataset).
- *Source prediction:* Along with predicting our Real and Fake, our model can be used to track the source of the corresponding images too.
- *Task-specific Adaptations:* Dual-head setups and *KANLinear* layers enable tailored solutions for complex scenarios and provides a promise against catastrophic forgetting.
- *Adversarial Training and Defense:* By using a variety of defense mechanisms during training and testing, and training on a combination of real and adversarial image data, our model demonstrates resistance to adversarial attacks.

4.2 Task-2

4.2.1 ***Dataset Preparation***. As there was no dataset available which had images with artifact annotations, we chose to create a fully annotated dataset on our own. For this, we followed a two step approach:

- First, randomly selected fake images belonging to all the stated models, and then used GPT-4o to extract artifacts and the reasoning behind those selected artifacts from the given list of artifacts using the prompt given below.
- Secondly, we used a Human in The Loop approach to manually check all the images and the generated responses from GPT-4o.

Table 6: Task 2 Fine-tuning Dataset Composition

Dataset	No. of Fake Images
BigGAN	450
Midjourney	825
Stable Diffusion v5	825
GigaGAN	38
StyleGAN	200
TOTAL	**2,338**

Afterwards, we used this dataset to fine-tune our model and effectively analyze the reasons behind fake images.

Figure 4: The horse image used for inference testing.

4.2.2 ***Vision Language Models (VLMs)***. For this task we tested multiple VLMs and evaluated their performance by making inferences on the horse image (Figure 4) provided in the problem statement. We take the output of the VLM along with the ground truth annotation given in the document and pass it to GPT-4o to get an estimated similarity score for every tested VLM. Table 7 shows the comparative analysis of all tested VLMs.

Table 7: Comparative analysis of all tested VLMs based on similarity scores.

VLM	Similarity Score (in %)	Remarks
InternVL	30-35	Low accuracy
Qwen	40-45	Low accuracy
Ovis	40-50	Moderate accuracy
LLaVA	**65-70**	**Best performer**

4.2.3 ***Knowledge Distillation***. Knowledge distillation is a technique that aims to transfer the learnings of a large pre-trained model, the "teacher model," to a smaller "student model". We aim to train a compact model (LLaVA (v1.6 Vicuna 7b)) to mimic a much larger model (GPT-4o with over 200B parameters).

We downscaled the images used in the annotated dataset and later used it to fine-tune our selected model from the previous experiments to perfectly fit the use-case of this task. Here GPT-4o works as a Teacher model whose outputs are used to fine-tune the LLaVA model, allowing the LLaVA model to achieve results which mimic the results of the large GPT-4o model, while maintaining a compact size.

Prompt used for data preparation

```
{
  You are an expert image analyzer excelling in analyzing fake images based on some parameters. This is an AI
     generated image. Please analyze it and return a JSON containing the artifact names and the reasons on the
     basis of this you would mark it fake. Return a JSON with keys as relevant artifact names. Do not make up
     any new artifact by yourself and be clear and confident with the reasons you have given. The artifacts to
     look for are as follows:
        - Inconsistent object boundaries
        - Discontinuous surfaces
        - Non-manifold geometries in rigid structures
        - Floating or disconnected components
        - Asymmetric features in naturally symmetric objects
        - Misaligned bilateral elements in animal faces
        - Irregular proportions in mechanical components
        - Texture bleeding between adjacent regions
        - Texture repetition patterns
        - Over-smoothing of natural text        (Rest of attributes)
}
```

4.2.4 *LLaVA (v1.6 Vicuna 7b)*.

LLaVA (v1.6 Vicuna 7b) (Language-Vision Alignment via Attention) focuses on aligning textual prompts with visual features using cross-modal attention mechanisms. Its ability to identify subtle visual patterns makes it highly suitable for artifact detection. By leveraging this model, we could pinpoint irregularities in images, such as unnatural textures or lighting inconsistencies, and generate textual explanations that align closely with the detected features.

This model stood out during our experiments for being not only compact, but also accurate. In our approach we fine-tune this model.

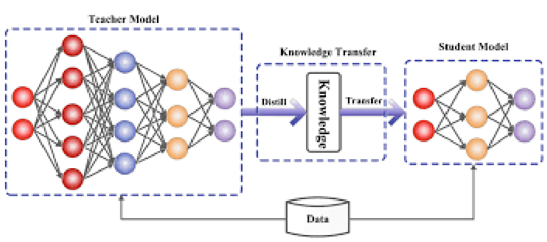

Figure 5: Knowledge Distillation

4.2.5 *Fine tuning*.

The fine-tuning was done with LoRA [8], a parameter-efficient fine-tuning method on the LLaVA v1.6 model, based on Vicuna 7B. LoRA adapts the model to become powerful while still preserving the pre-trained model. The salient features of the above approach are as follows:

Table 8: Fine-Tuning Configuration

Parameter	Value
GPUs Used	3 (A30, 24GB)
Train Batch Size (per device)	4
Gradient Accumulation Steps	1
Warm-up Rate	0.03
Epochs	3
Learning Rate	2e-4
Optimizer	adamw
LoRA_r	128
LoRA Alpha	256
LR Scheduler Type	Cosine
Model Max Length	2048

(1) **Efficiency:** LoRA minimizes computational and memory requirements by focusing on small, trainable rank-decomposition matrices injected into each layer of the model. This significantly reduces the need to update all parameters, thus allowing for fine-tuning of large models like LLaVA 1.6 (7B) on resource-constrained hardware.
(2) **Modularity:** LoRA is designed in a modular fashion, and the original model weights are not modified. LoRA modules can be fine-tuned independently for a specific task and loaded when necessary, which makes it very flexible and easy to work with different applications.
(3) **Preservation of Pretrained Knowledge:** Given that the base model weights are largely frozen, fine-tuning retains the core knowledge of the original LLaVA v1.6 Vicuna 7B model. This approach ensures that the model adapts effectively to new tasks without losing its foundational capabilities.

4.2.6 *Advantages of the Proposed Methodology*.

The proposed methodology combines dataset distillation, high-resolution inference, human-in-the-loop validation, and LoRA-based fine-tuning to achieve a cost-efficient and performance-oriented approach. Key advantages include:

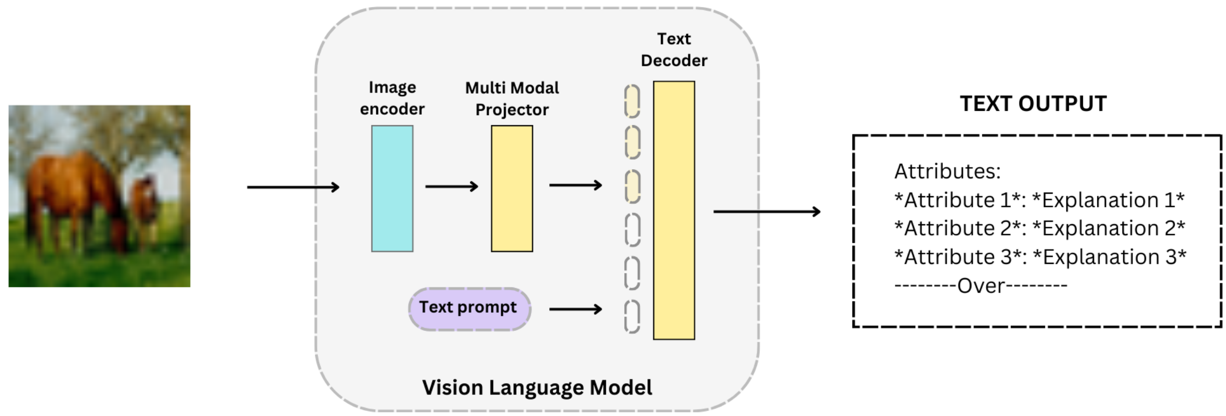

Figure 6: High-resolution (first row) versus low (32×32 pixel) resolution (second row) images.

- **Cost and Time Efficiency:** The integration of dataset distillation minimizes the manual effort required for data annotation, significantly reducing both time and financial costs. The modularity of LoRA further optimizes computational requirements, enabling fine-tuning on resource-constrained hardware.
- **Improved Model Generalization:** High-resolution inference and robust handling of different resolutions ensure that the model generalizes effectively across diverse practical scenarios, maintaining accuracy and adaptability.
- **Enhanced Explainability:** Vision-Language Models inherently provide interpretable outputs by combining visual and textual reasoning. Additionally, human oversight ensures that outputs align with human intuition, improving trust in the model's predictions.
- **Dynamic Learning Capabilities:** The human-in-the-loop validation mechanism fosters continuous learning by allowing dynamic updates to the dataset and iterative refinement of the model. This supports long-term relevance and performance improvement.
- **Scalability and Flexibility:** The proposed approach is highly scalable, with LoRA enabling efficient adaptation to new tasks and domains. This flexibility makes it well-suited for a variety of applications, ranging from edge device deployment to large-scale model training.
- **Preservation of Foundational Knowledge:** LoRA's parameter-efficient design preserves the pre-trained knowledge of the foundational model (e.g., LLaVA v1.6 Vicuna 7B), ensuring that fine-tuning adapts the model to new tasks without erasing its existing capabilities.
- **Seamless Workflow Integration:** The methodology combines all steps—dataset creation, inference, validation, and fine-tuning—into a streamlined workflow. This holistic integration reduces iterative development cycles, enabling rapid experimentation and deployment.
- **Resource Optimization:** By leveraging lightweight fine-tuning techniques such as LoRA, the methodology supports deployment on devices with limited computational resources, including edge devices, while still maintaining performance.

This comprehensive set of advantages positions the proposed methodology as a robust and versatile framework for efficiently adapting Vision-Language Models to high-resolution tasks in a scalable and explainable manner.

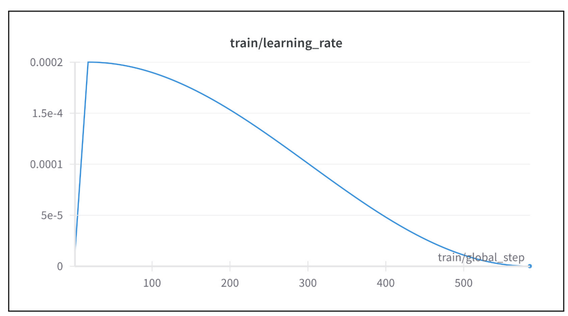

Figure 7: Learning Rate vs Step

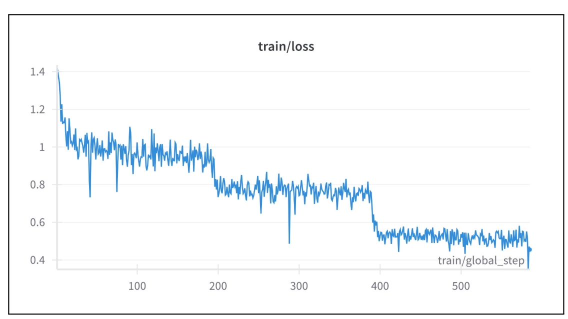

Figure 8: Loss vs Step

5 Current Limitations

Some of the limitations faced by us and present in the proposed solution are as follows:

- **Model Size and Feature Extraction:** A major limitation we faced was the limit on the size of the model, making smaller model demanded a trade-off between accuracy and the quality of reasons getting generated. Although this led to reduced inference time, the models faced difficulty in learning efficient and richer embedding spaces.

- **Limited Ability to Use Few-Shot Learning:** The VLMs used had less number of parameters, which prevented us from employing few-shot learning [16] as the model had limited context window. This reduces the model's ability to generalize well across unseen data. The small context window made it difficult to capture the broader relationships between textual and visual elements, leading to poorer accuracy in reasoning.

- **Artifacts in the Dataset:** We were also constrained by the fact that the list of artifacts was limited to a number of artifacts, some of the images within the dataset may have had artifacts that were not similar to the ones in the list, making it difficult to reason about their presence and relevance. These artifacts could be inconsistencies or noise in the data, such as improper text formatting or incorrect background elements, which hindered the model's ability to focus on meaningful features. The limit on the artifacts may have limited its ability to recognize complex visual patterns or text-image correlations.

- **Limited Computational Resources:** Another limitation was the limited computational power, which prevented us from training the reasoning model (VLM) on the full dataset of 300,000 images. Due to the resource constraints, we had to use a subsample of only 2,338 images. These images were carefully selected to represent the key features and attributes, but the reduced dataset size limited the model's exposure to the full diversity of images that would have been beneficial for training. This limitation could have significantly affected the model's ability to achieve the highest possible performance on real-world data.

- **Adversarial Training Trade-Off:** Increasing the percentage of adversarial images during training resulted in a trade-off for the model's accuracy. Specifically, when adversarial images were included, the model started predicting more images as fake, even though they were natural. This highlighted the difficulty in balancing adversarial training—intended to make the model more robust—while maintaining its accuracy in distinguishing genuine data from fake data. This trade-off underlined the challenges in achieving a stable model performance when adversarial examples are heavily introduced in the training process. We chose to keep the model trained on low percentage of adversarial examples in training data because it led to an increase in 5-6% accuracy increase for natural images while reducing the accuracy on fake images by 1%.

- **Lack of Feature Localization:** Additionally, the model was incapable of localizing specific features or attributes in an image, which made it difficult to interpret and understand the visual components in detail. Without this ability, the model struggled probably to make precise associations between visual elements and corresponding textual descriptions. This limitation hindered its performance in tasks that required detailed feature extraction, such as identifying specific objects or understanding nuanced text-image interactions.

6 Possible Improvements and Future Scope

The future scope of the proposed solution and some possible improvements can be extended in the following directions:

- **Human Annotated Dataset:** A dataset containing ample amount of human annotated or human verified (through our Human in The Loop approach) containing images and their artifacts can be created. This dataset will allow us to expose our model to more diverse and high quality training data, allowing it to generalize better across different cases and give a better predictive performance. Additionally, if we can create a dataset which has localized annotations, it can learn to recognize specific features and patterns within distinct regions.

- **Classification Model (Task 1) Distillation:** The current best model has around 12 M parameters. To optimize it for use on mobile devices, we can distill the knowledge from our current best model to a smaller backbones using the KAN layer embeddings. After this we can directly use the pre-trained MLP classifiers for classification.

- **Icorporating Richer Embeddings:** Using contrastive loss or triplet loss based training between the embeddings of the two classification heads, we can create an embedding space where the representations formed are richer in their distribution. This can help enhance the classification tasks.

- **Fine-Tuning VLM on Real vs Fake Classification:** We can fine-tune the Vision encoder of the Vision Language Model (VLM) for real vs fake detection by training it on a large dataset (300K+ examples). This pre-training will possibly allow the model to learn how to differentiate between real and fake. Fine-tuning on a smaller annotated dataset after the above pre-training can help boost the models reasoning capabilities and give enhanced results on artifact identification and reasoning task.

- **Few Shot Prompting:** The small context window of the current VLM makes it impossible to achieve good results with few shot prompting. Using larger models with broader context windows could improve few-shot prompting capabilities, allowing the model to generalize better from limited data, making it more adaptable to new tasks.

Prompt used for inferencing

```
{
  "image_file": "horse.png",
  "prompt": "The given image is AI generated. Give me reasons why it is AI generated. You can select from the
     following list of artifacts:
        - Inconsistent object boundaries
        - Discontinuous surfaces
        - Non-manifold geometries in rigid structures
        - Floating or disconnected components
        - Asymmetric features in naturally symmetric objects
        - Misaligned bilateral elements in animal faces
        - Irregular proportions in mechanical components
        - Texture bleeding between adjacent regions
        - Texture repetition patterns
        - Over-smoothing of natural textures
        - Artificial noise patterns in uniform surfaces .....(rest of attributes)""
        }
```

Example inference output

```
{
    Output {
    "Inconsistent Object Boundaries": "The edges of the horse blend unnaturally with the background, indicating
       inconsistent boundaries.",
    "Discontinuous Surfaces": "The transition between the horse's neck and body appears disconnected.",
    "Asymmetric Features in Naturally Symmetric Objects": "The horse's legs and body exhibit asymmetry that is
       not typical for real horses.",
    "Improper Fur Direction Flows": "The fur or mane on the horse does not follow a natural flow pattern.",
    "Inconsistent Shadow Directions": "The shadows cast by the horse do not align with a single light source.",
    "Multiple Light Source Conflicts": "There are conflicting highlights and shadows suggesting multiple light
       sources.",
    "Unnatural Pose Artifacts": "The pose of the horse seems unnaturally rigid and staged.",
    "Artificial Smoothness": "The texture of the horse's body appears overly smooth and lacks natural detail." }
}
```

7 Supplementary Methods

Apart from our main approach, we employed some other approaches that didn't result in our optimal choice for the classification task. Below are brief descriptions of our experiments.

(1) **Training a CNN Model on Real and Fake Labels**
We tried to fit our dataset, which consists of images from BigGAN, StyleGAN, Stable Diffusion, Midjourney, etc., using a traditional CNN model (e.g., ResNet-50, ResNet-18, MobileNet, ResNet-12, etc.). However, this approach resulted in poor training and validation accuracy compared to our finalized method. The following table provides a comparative view of the results.

Model	Accuracy
ResNet-12	37.16%
ResNet-18	58.23%
ResNet-50	63.19%
EfficientNet B0	59.64%
MobileNet	62.57%
EfficientViT	62.38%

Table 9: Real vs Fake Classificiation Accuracies.

(2) **Multi-Class Classification-Based Approach**
We also tried a multi-class classification model using a traditional CNN backbone for feature extraction to classify images as real natural images, Generative Adversarial Network (GAN)-generated, and diffusion-generated content. This approach yielded less favorable results compared to our finalized method. The following table summarizes the results.

Achieved accuracy was 78% using an ensemble of three ResNet-50 classifiers.

(3) **Ensemble Method**
We also experimented with a classical ensemble method to improve results through voting. Each model was trained independently on diffusion, GAN and midjourney based images independently from our mentioned dataset. This method produced better results compared to the other approaches mentioned above, but it became computationally intensive as the number of parameters increased, leading to a rise in inference time.

Model	Accuracy
EfficientNet B0	51.85%
EfficientNet B3	54.27%
ResNet50	48.81%

Table 10: Multiclass classification Accuracies.

(4) **Randomization of Input Layer**
We experimented with [23] randomization of the input layer using various operations like random resizing with random zero padding in a random manner. With the purpose as a defense against adversarial attacks both single-step and interactive attacks. While testing, accuracy against attacks increased but it decreased for non adversarial images. We don't use this method because it requires further experimentation and research, but it has proven effective against adversarial attacks.

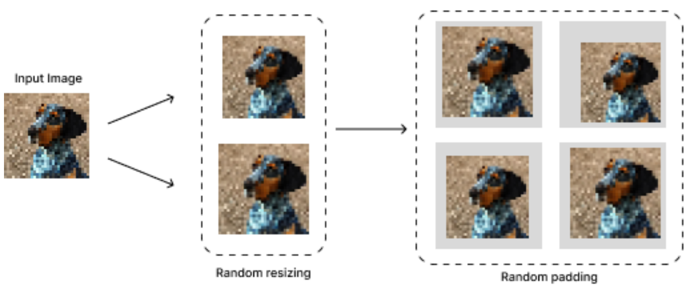

Figure 9: Randomization of Input layer

(5) **Few-Shot Prompting**
We also tried to incorporate few-shot reasoning mechanisms such as prototype-based methods and fine-tuning on small support sets into our approach to enhance the identification of artifacts and similarity scoring. These methods, though very promising in improving the reasoning of accuracy for novel artifacts with fewer examples, introduced inconsistencies in scoring similar artifacts that are well represented. Due to further refinement and balancing generalization with specificity of reasoning, these methods were not adopted in their present form despite their merit for few-shot reasoning scenarios.

(6) **Multi-Label Classification**
We investigated a Multi-Label Classification approach to identify the artifacts present in a given input image with the intent of using such predictions as inputs to a Vision-Language Model for more accurate reasoning. While this sounded promising for improving the reasoning pipeline by providing detailed context at the artifact level, the primary challenge was that the availability of annotated multi-label datasets was quite limited and hence hard for the model to generalize effectively across diverse scenarios. More research and enlargement of the dataset would be required to realize its full potential.

8 Use Cases and Applications

(1) **AI-Generated Fake Product Reviews in E-Commerce**
- **Use Case:** Fake product reviews accompanied by AI-generated images are becoming a serious problem in e-commerce. These reviews often feature low-resolution, seemingly authentic "product shots" that mislead buyers and do not match the actual products being sold.

(2) **Synthetic Image Detection in Automated Video Content Moderation**
- **Use Case:** On video sharing platforms like youtube, instagram, users may use AI-generated image as part of video thumbnail and content. These synthetic images could be used to mislead viewers about the video's content or to evade content detection systems.
- **Application:** These video platforms can implement AI systems to detect such low resolution synthetic images in video thumbnails. By analyzing patterns and pixel-level artifacts, these models can flag videos which tries to misleading users before they are shared with millions of viewer, improving trust in the platform.

(3) **Low-Resolution Profile Picture Fraud in Online Dating Apps**
- **Use Case:** To impersonate others and deceive potential matches or to engage in fraudulent activity such as scamming and catfishing, users may upload low-resolution AI-generated profile pictures.
- **Application:** Dating platforms like Tinder or Bumble can use AI-powered systems to detect low-resolution synthetic profile images automatically and flag them as suspicious and thus preventing users to use them. By identifying unnatural patterns or pixel-level artifacts, the platform can prevent fake profiles from deceiving users, making the app more authentic for its community.

9 Conclusion

Our solution successfully tackles the pressing challenge of detecting AI-generated images in an era of rapid advancements in generative technologies such as GANs, diffusion models, and Stable Diffusion. The proposed framework, centered around a multi-head CNN architecture and fine-tuned Vision-Language Models (VLMs), showcases significant potential for addressing the complexities of low-resolution image classification and artifact detection.

By introducing innovations such as KANLinear layers, adversarial training, and modular architectures, the methodology effectively enhances model performance while maintaining computational efficiency. These innovations enable the detection system to adapt to a wide variety of generative techniques, ensuring high accuracy across datasets with diverse image characteristics. Moreover, the use of knowledge distillation to fine-tune compact VLMs aligns with the growing need for scalable and resource-efficient solutions, making this approach suitable for deployment in rdeal-world applications.

In addition to achieving robust classification of real versus AI-generated images, the framework's ability to identify specific artifacts and provide interpretable explanations marks a significant step toward improving transparency in AI-powered detection systems. The integration of human-in-the-loop validation ensures that the system's outputs align with human intuition, thereby fostering trust in its predictions. This combination of interpretability and performance positions the solution as a reliable tool for use in digital forensics, e-commerce fraud prevention, and online content moderation.

Despite its strengths, our solution highlights a few limitations that provide avenues for future research. Constraints such as limited access to annotated datasets, the small size of the fine-tuned VLM, and computational resource limitations have restricted the system's ability to generalize across all possible scenarios. These challenges emphasize the importance of expanding datasets, exploring localized artifact detection, and developing lightweight models capable of handling large-scale, diverse data.

The use cases, including detecting fraudulent e-commerce reviews, synthetic video thumbnails, and fake profile images, demonstrate the potential societal impact of this research. By addressing these real-world problems, the system contributes to creating safer and more trustworthy digital ecosystems.

In conclusion, this work not only addresses a critical technological challenge but also provides a strong foundation for future advancements in the detection of AI-generated content. By refining and extending this methodology, it is possible to build even more robust systems that keep pace with the rapid evolution of generative models, ensuring continued trust and authenticity in digital media.

References

[1] Jordan J. Bird and Ahmad Lotfi. 2024. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access* 12 (2024), 15642–15650. https://doi.org/10.1109/ACCESS.2024.3356122
[2] Ali Borji. 2023. Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2. arXiv:2210.00586 [cs.CV] https://arxiv.org/abs/2210.00586
[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv:1809.11096 [cs.LG] https://arxiv.org/abs/1809.11096
[4] X. Chai et al. 2020. Detecting deepfake images with generalizable classifiers. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.
[5] S. Chhabra et al. 2023. Methods for enhancing sub-artifacts in deepfake detection. *Journal of Machine Learning* (2023).
[6] Yunfeng Diao, Naixin Zhai, Changtao Miao, Xun Yang, and Meng Wang. 2024. Vulnerabilities in AI-generated Image Detection: The Challenge of Adversarial Attacks. arXiv:2407.20836 [cs.CV] https://arxiv.org/abs/2407.20836
[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML] https://arxiv.org/abs/1503.02531
[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] https://arxiv.org/abs/2106.09685
[9] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up GANs for Text-to-Image Synthesis. arXiv:2303.05511 [cs.CV] https://arxiv.org/abs/2303.05511
[10] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv:1812.04948 [cs.NE] https://arxiv.org/abs/1812.04948
[11] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML] https://arxiv.org/abs/1312.6114
[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV] https://arxiv.org/abs/2304.08485
[13] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. 2024. KAN: Kolmogorov-Arnold Networks. arXiv:2404.19756 [cs.LG] https://arxiv.org/abs/2404.19756
[14] P. Lorenz, R. L. Durall, and J. Keuper. 2023. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *International Conference on Computer Vision*. 448–459.
[15] A. Ojha et al. 2023. The non-generalizability of neural networks to unknown families of generative models. *arXiv preprint* (2023).
[16] Archit Parnami and Minwoo Lee. 2022. Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. arXiv:2203.04291 [cs.LG] https://arxiv.org/abs/2203.04291
[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752
[18] A. B. Shiyam and G. Poravi. 2023. Deepfake low resource image detection with explainable reporting. In *Proceedings of [Conference Name]*.
[19] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs.LG] https://arxiv.org/abs/1905.11946
[20] Hetvi Waghela, Jaydip Sen, and Sneha Rakshit. 2024. Robust Image Classification: Defensive Strategies against FGSM and PGD Adversarial Attacks. arXiv:2408.13274 [cs.CR] https://arxiv.org/abs/2408.13274
[21] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. YOLOv10: Real-Time End-to-End Object Detection. arXiv:2405.14458 [cs.CV] https://arxiv.org/abs/2405.14458
[22] S.-Y. Wang et al. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Conference on Computer Vision and Pattern Recognition*. 8695–8704.
[23] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. 2018. Mitigating Adversarial Effects Through Randomization. arXiv:1711.01991 [cs.CV] https://arxiv.org/abs/1711.01991

Appendix

A Few Shot Prompt

Few Shot Prompt

```
{
USER: <image>
The given image is fake, can you give me reasons for why it is fake, listing recognisable attributes and their
    explanations
ASSISTANT:
Inconsistent object boundaries: The boundary of the horse against the background appears slightly blurred,
    particularly near the legs.\nTexture bleeding between adjacent regions: Noticeable blending of textures,
    especially between the horses body and its shadow.\nArtificial smoothness: The horses body exhibits a
    smoothness lacking natural texture.\nIncorrect shadow directions: The shadow underneath the horse does not
    align with the light source.\nUnrealistic eye reflections: The eye reflection appears overly pronounced.\
    nOver-smoothing of natural textures: The horse lacks detailed fur patterns typically present in natural
    images.\nMultiple inconsistent shadow sources: Shadows originate from conflicting directions.\nUnnatural
    lighting gradients: Lighting transitions are too smooth and unnatural.
USER: <image>
The given image is fake, can you give me reasons for why it is fake, listing recognisable attributes and their
    explanations
ASSISTANT:
Inconsistent object boundaries: The edges of the objects in the image are not clearly defined, suggesting a
    blending of separate elements.\nTexture bleeding between adjacent regions: The texture of the yellow object
    bleeds into the background, showing inconsistency.\nArtificial noise patterns in uniform surfaces: There is
    visible noise across the uniform areas that looks artificially generated.\nUnnatural color transitions: There
     are abrupt color changes between objects and the background that do not appear natural.\nIncorrect
    reflection mapping: Reflection on surfaces seems incorrectly mapped, not aligning with realistic lighting and
     perspectives.
USER: <image>
The given image is AI generated. Give me reasons why it is AI generated. You can select from the following list
    of artifacts:\n- Inconsistent object boundaries\n- Discontinuous surfaces\n- Non-manifold geometries in rigid
     structures\n- Floating or disconnected components\n- Asymmetric features in naturally symmetric objects\n-
    Misaligned bilateral elements in animal faces\n- Irregular proportions in mechanical components\n- Texture
    bleeding between adjacent regions\n- Texture repetition patterns\n- Over-smoothing of natural textures\n-
    Artificial noise patterns in uniform surfaces\n- Unrealistic specular highlights\n- Inconsistent material
    properties\n- Metallic surface artifa    (rest of attributes)
ASSISTANT:
}
```

B Artifacts Analysis

Table 11: Artifacts Observed in AI-Generated Horse Images

Model	Artifacts Observed
LLaVA v1.6 Vicuna 7b (Non Finetune)	• **Inconsistent object boundaries:** The boundary of the horse against the background appears slightly blurred, particularly near the legs. • **Texture bleeding between adjacent regions:** Noticeable blending of textures, especially between the horse's body and its shadow. • **Artificial smoothness:** The horse's body exhibits a smoothness lacking natural texture. • **Incorrect shadow directions:** The shadow underneath the horse does not align with the light source. • **Unrealistic eye reflections:** The eye reflection appears overly pronounced. • **Over-smoothing of natural textures:** The horse lacks detailed fur patterns typically present in natural images. • **Multiple inconsistent shadow sources:** Shadows originate from conflicting directions. • **Unnatural lighting gradients:** Lighting transitions are too smooth and unnatural.
Qwen (40-45%)	• **Inconsistent object boundaries:** The horse's body appears made of disconnected parts. • **Discontinuous surfaces:** The horse's body shows jagged edges and unnatural transitions. • **Non-manifold geometries:** Legs and body appear as disconnected parts. • **Floating components:** The horse's legs and body appear to float mid-air rather than being grounded. • **Asymmetric features:** The horse's face appears overly symmetric, which is unnatural. • **Unrealistic specular highlights:** The eyes reflect light unrealistically. • **Improper fur direction flows:** The horse's fur flows unnaturally. • **Misshapen ears:** Ears appear anatomically incorrect. • **Scale inconsistencies:** Body parts are disproportionate. • **Inconsistent shadow directions:** Shadows appear from unexpected directions. • **Multiple light source conflicts:** Shadows originate from multiple sources.
Ovis1.6-Gemma2-9B	• **Blurred boundaries:** The horse's edges lack sharp definition. • **Texture repetition:** The coat texture appears repetitive, lacking natural variation. • **Over-smoothing:** The coat and mane are overly smooth. • **Unrealistic specular highlights:** Highlights are inconsistent with natural lighting. • **Improper fur direction:** The fur flows unnaturally. • **Unrealistic eye reflections:** Eyes lack depth and realism. • **Anatomically incorrect features:** The paws appear disproportionate. • **Unnatural pose:** The pose lacks natural balance. • **Scale inconsistencies:** Body parts are not proportionally accurate.
LLaVA - v1.6 Vicuna 7b(Fine-tuned)	• **Inconsistent object boundaries:** Edges blend unnaturally with the background. • **Discontinuous surfaces:** Transition between the neck and body is disconnected. • **Asymmetric features:** Legs and body show unnatural asymmetry. • **Improper fur direction flows:** Fur and mane do not follow natural patterns. • **Inconsistent shadow directions:** Shadows are misaligned with a single light source. • **Multiple light source conflicts:** Highlights and shadows suggest multiple light sources. • **Unnatural pose:** The pose appears rigid and staged. • **Artificial smoothness:** The body texture lacks natural detail.