INTER IIT
TECH MEET 13.0

Adobe

IMAGE CLASSIFICATION AND ARTIFACT DETECTION

TEAM 57

AGENDA

AI Generated Image of Truck from Adobe Firefly

Real Image of Truck

Task 1

Detection of AI Generated Images

Development of a robust model that can accurately identify whether an image(with Adversarial Perturbations) is AI-generated or Real.

APPROACH FOR TASK 1

The proposed architecture is as follows:

- ResNet50[1] inspired backbone for robust feature extraction in spatial domain.
- Wavelet attention, and Freqnet to capture frequency information.
- Wavelet Attention Mechanism captures high and low-frequency information and filters out noise.
- FreqNet captures high-frequency representations of images and features.

WAVELET ATTENTION

- Discrete wavelet transforms are used in the wavelet attention(WA) block[3].
- DWT2D class is used from the pytorch-wavelets library. The wavelet family used is ‘haar’.
- DWT gives us output low-frequency component X_{ll} and three high-frequency components X_{lh}, X_{hl}, and X_{hh}.
- We discard the X_{hh} as it contains the maximum amount of noise.

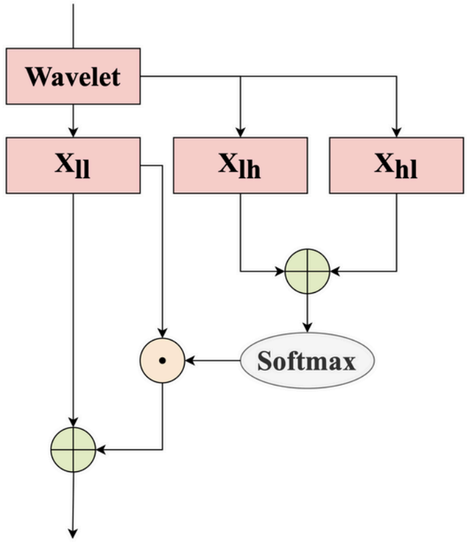


Figure: Wavelet Attention Block

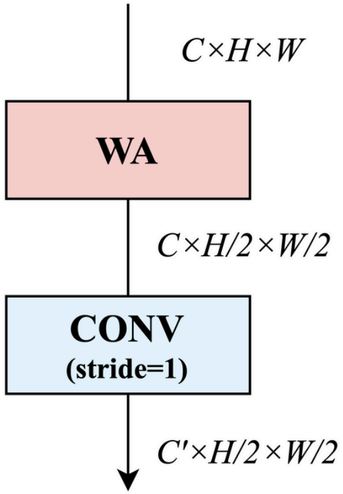


Figure: Wavelet Attention Stride

FREQNET

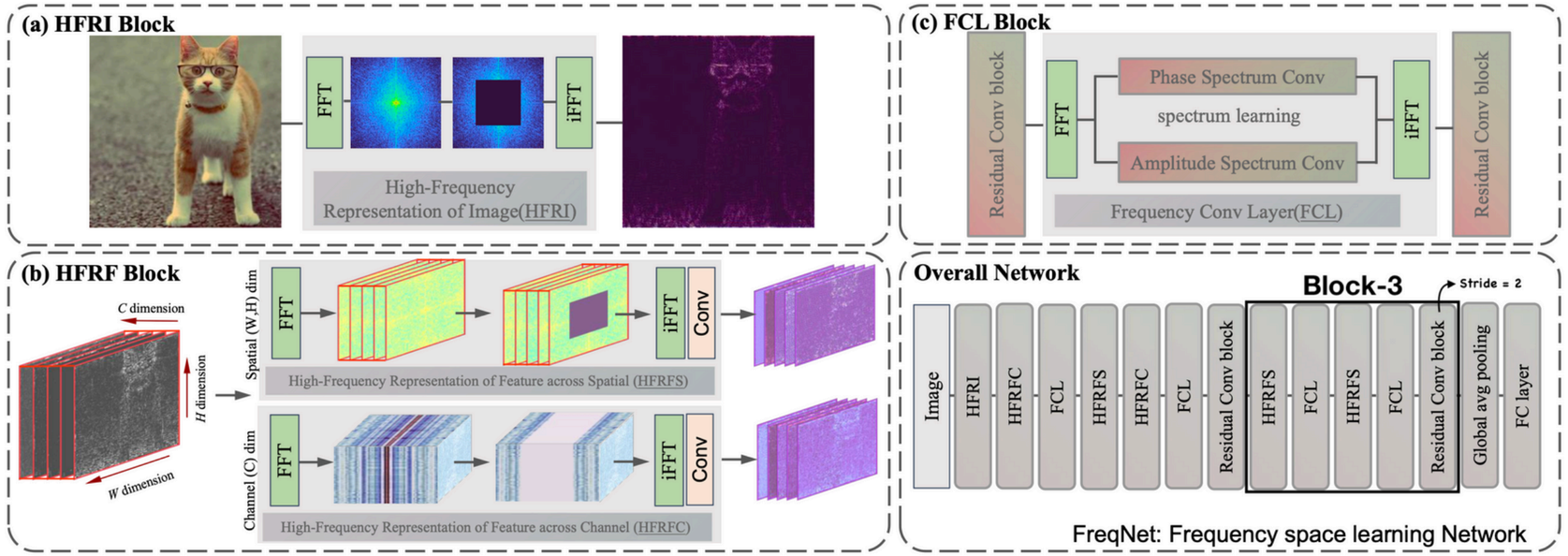
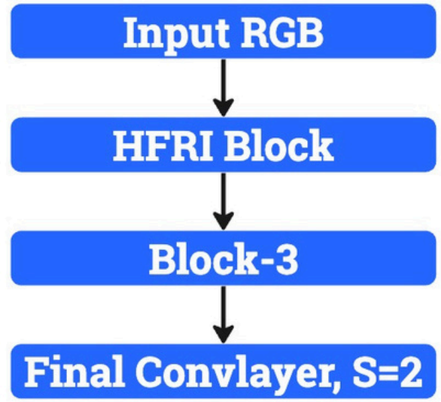
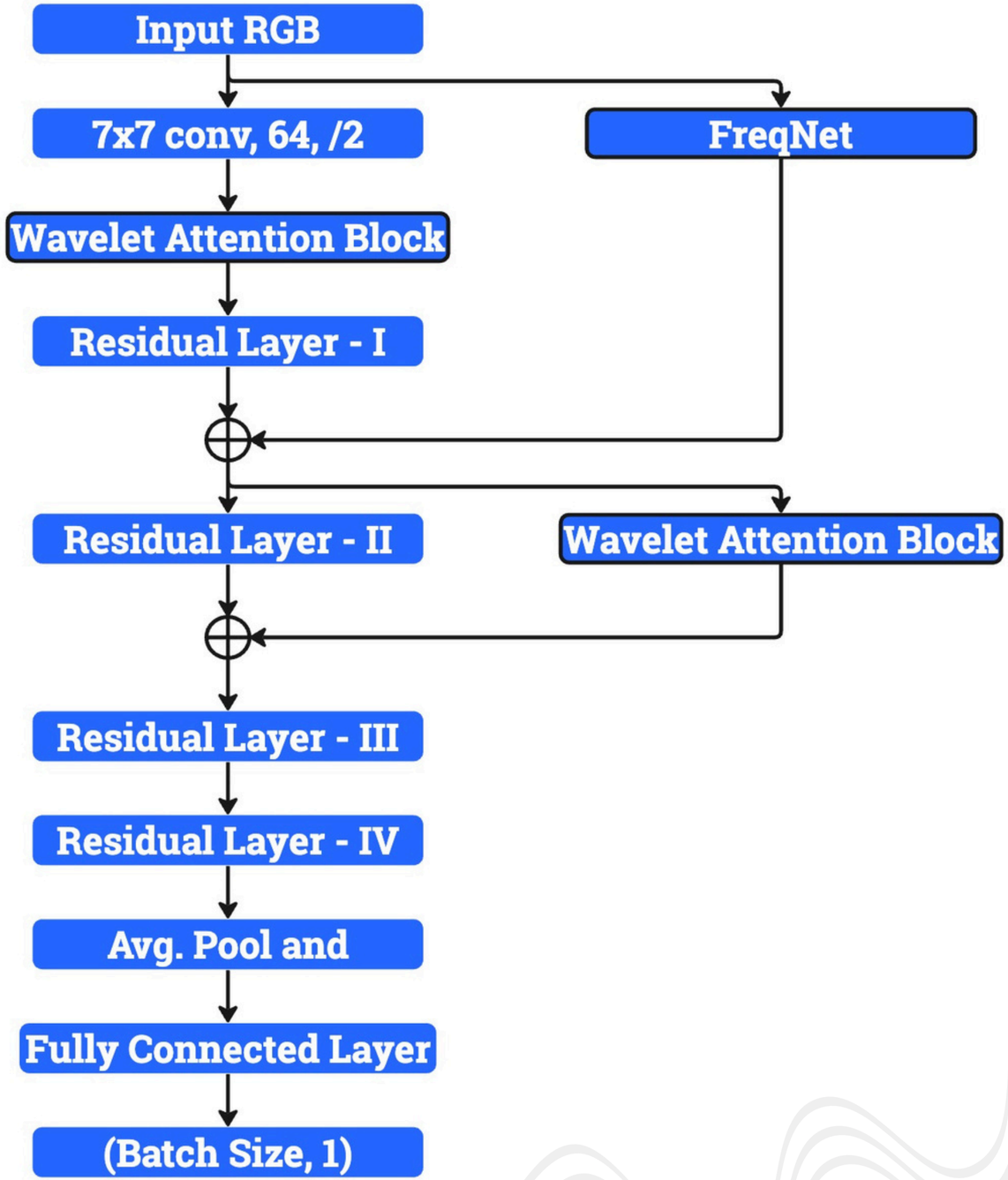


Figure: Freqnet Architecture

- Converting an image to the frequency domain using FFT and iFFT.
- Only one block of the original FreqNet[2] is used along with the HFRI block.
- The residual convolution layer and the final Convlayer have stride as 2.



PIPELINE FOR TASK 1



TRAINING PROCEDURE & RESULTS

EPOCHS	**100**
INITIAL LEARNING RATE	5e-3
OPTIMIZER	**Adam**
LOSS FUNCTION	BCEWithLogitsLoss
BATCH SIZE	640
LEARNING RATE SCHEDULER	ReduceLROnPlateau
PATIENCE	5
NUMBER OF PARAMETERS	24.8M
RESULTS	**TRAINING ACC.** : 97.44% **VALIDATION ACC**: 95.98%

Task 2

Artifact Identification & Explanation

Development of a model that can effectively identify and provide explanations of all the Artifacts that are present in an image.


```
"Artifacts annotation": [
    {
      "rect_start": [87,272],
      "rect_end": [215,431],
      "artifacts_caption": "Wizard's fingers are distorted."
      "artifacts_class": "Distorted and Deformated Components"
    }
]

"Other artifacts caption": "None"
```


SynArtifact: Classifying and Alleviating Artifacts in Synthetic Images via Vision-Language Model[1]

APPROACH FOR TASK 2

The proposed solution involves two vision language models.

1. CLIP (Contrastive Language Image Pre-training)
2. Idefics 2

CLIP was used for Artifact identification whereas Idefics 2 is used for artifact explanation.

Both the VLMs used were pretrained.

APPROACH FOR TASK 2

VISION LANGUAGE MODELS USED FOR TASK 2	CLIP - openai/clip-vit-base-patch32 Idefics2 - HuggingFaceM4/idefics2-8b
CLIP[4]	No. of artifacts used for detection - 34(filtered based on image size and classes of image) Threshold - 0.05 Text encoder parameters - 86M Image encoder parameters - 115M
IDEFICS2[5]	Model: Idefics2-8b (Conditional Generation) Quantization: 4-bit (NF4 format) Compute Dtype: torch.float16

PIPELINE FOR TASK 2

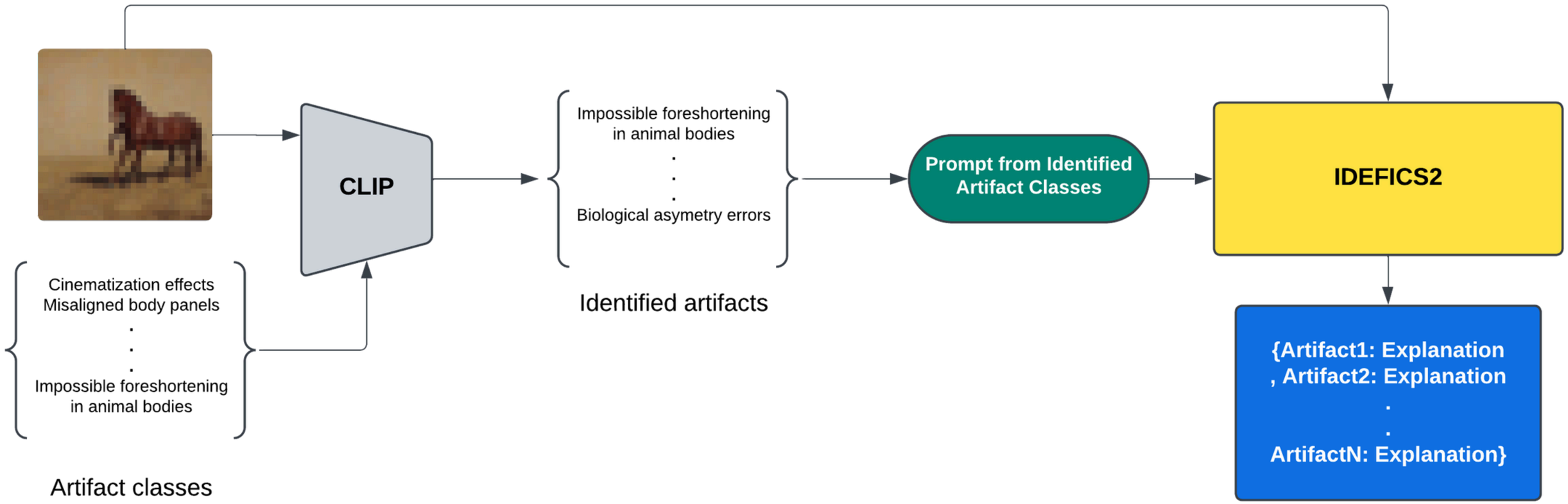


Figure: Pipeline for Task 2: The CLIP identifies artifact classes which are fed into Idefics2 for explanation

Note: The model used is not End-to-End.

RESULTS

Figure: one of the images of testing dataset

CLIP


```
{
  "index": 88,
  "explanation": [
    "Cinematization Effects",
    "Movie-poster like composition of ordinary scenes",
    "Exaggerated characteristic features",
    "Ghosting effects: Semi-transparent duplicates of elements",
    "Unrealistic specular highlights",
    "Aliasing along high-contrast edges",
    "Artificial enhancement artifacts"
  ]
}
```


Example of one image used to detect artifacts using CLIP.
the results can also be visually cross verified.

RESULTS

Figure: one of the images of testing dataset


```
"index": 88,
"explanation": [
  "Cinematization Effects",
  "Movie-poster like composition of ordinary scenes",
  "Exaggerated characteristic features",
   "Ghosting effects: Semi-transparent duplicates of elements",
  "Unrealistic specular highlights",
  "Aliasing along high-contrast edges",
  "Artificial enhancement artifacts"
]
}
```


Artifacts detected from CLIP

A descriptive prompt consisting of artifacts

Idefics2


```
{
  "index": 88,
  "explanation": {
      "Cinematization Effects": " The image has a cinematic effect, with the colors and lighting being exaggerated",
     "Movie-poster like composition of ordinary scenes": " The image has a movie poster-like composition, with the people and objects being arranged in a way that is reminiscent of a movie poster",
      "Exaggerated characteristic features": " The people and objects in the image have exaggerated characteristic features, such as larger-than-life eyes or exaggerated facial expressions",
    "Ghosting effects": " There are semi-transparent duplicates of elements in the image, such as people or objects, that give the image a ghosting effect",
    "Unrealistic specular highlights": " The specular highlights in the image are unrealistic, with bright spots that are not consistent with the lighting and environment",
    "Aliasing along high-contrast edges": " There is aliasing along high-contrast edges, such as the edges of people or objects, which causes the image to look unrealistic and jagged",
     "Artificial enhancement artifacts": " There are artificial enhancement artifacts in the image, such as over-sharpening or over-saturation, which make the image look unnatural and unrealistic"

  }
 }
```

MODIFICATIONS


```
{
    "The image contains the following artifacts": "\n\n1",
    "Aliasing along high-contrast edges": " This artifact is visible in the image as a jagged edge around the airplane",
     "The edges are not smooth and have a staircase effect": " The edges are not smooth and have a staircase effect",
      "This is due to the low resolution of the image": " This is due to the low resolution of the image",
    "2": "\n\n2",
     "Discontinuous surfaces": " This artifact is visible in the image as a discontinuous surface around the airplane",
      "The surface is not smooth and has a discontinuous appearance": " The surface is not smooth and has a discontinuous appearance",
    "3": "\n\n3",
       "Scale inconsistencies within single object": " This artifact is visible in the image as a difference in scale between the airplane and the other objects in the image",
       "The airplane is larger than the other objects, and the scale is not consistent": " The airplane is larger than the other objects, and the scale is not consistent",
      "These artifacts are important in terms of the image's overall composition, realism, and visual quality": "\n\nThese artifacts are important in terms of the image's overall composition, realism, and visual quality",
     "The image appears unrealistic and low-quality due to the presence of these artifacts": " The image appears unrealistic and low-quality due to the presence of these artifacts",
    "": ""
  }
```



```
{

    "Aliasing along high-contrast edges": " This artifact is visible in the image as a jagged edge around the airplane",
     "The edges are not smooth and have a staircase effect": " The edges are not smooth and have a staircase effect",
      "This is due to the low resolution of the image": " This is due to the low resolution of the image",
     "Discontinuous surfaces": " This artifact is visible in the image as a discontinuous surface around the airplane",
      "The surface is not smooth and has a discontinuous appearance": " The surface is not smooth and has a discontinuous appearance",
       "Scale inconsistencies within single object": " This artifact is visible in the image as a difference in scale between the airplane and the other objects in the image",
       "The airplane is larger than the other objects, and the scale is not consistent": " The airplane is larger than the other objects, and the scale is not consistent",
      "These artifacts are important in terms of the image's overall composition, realism, and visual quality": "These artifacts are important in terms of the image's overall composition, realism, and visual quality",
     "The image appears unrealistic and low-quality due to the presence of these artifacts": " The image appears unrealistic and low-quality due to the presence of these artifacts"

  }
```

DATASET PREPARATION

GENERATING FAKE IMAGES

A total of 2045 images were generated using the following generating models:

- Stable Diffusion: 1,500 images
- Adobe Firefly: 257 images
- Midjourney: 288 images

Image generated using stable diffusion

ANNOTATION

From the generated images, 637 were manually annotated with ChatGPT to identify artifacts, and the results were stored in JSON files named after each image.


```
{"Artifacts annotation": [
  {
    "artifacts_caption": "The aircraft has an unrealistic wing shape and alignment.",
    "artifacts_class": "18"
  },
  {
    "artifacts_caption": "The landing gear appears unsupported and not adequately detailed.",
    "artifacts_class": "22"
  },
  {
    "artifacts_caption": "The color patterns on the tail fin are oddly distributed.",
    "artifacts_class": "40"
  },
  {
    "artifacts_caption": "The front of the aircraft has an unnatural geometric shape.",
    "artifacts_class": "19"
  },
  {
    "artifacts_caption": "The overall perspective of the aircraft is skewed and inconsistent.",
    "artifacts_class": "62"
  }
 ]
}
```


Annotations generated using gpt 4o.

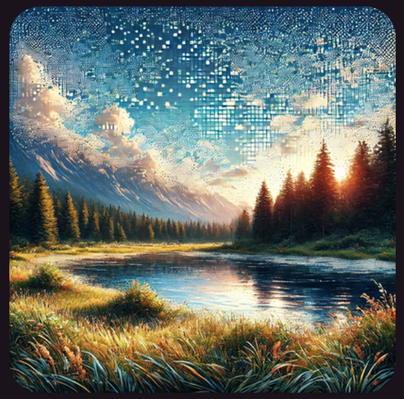
Key Challenges

Image Size

The images that were tesed for evaluation of model were of size 32x32.

Unavailability of dataset

Lack of appropriate dataset for explanations of artifacts in AI generated images.



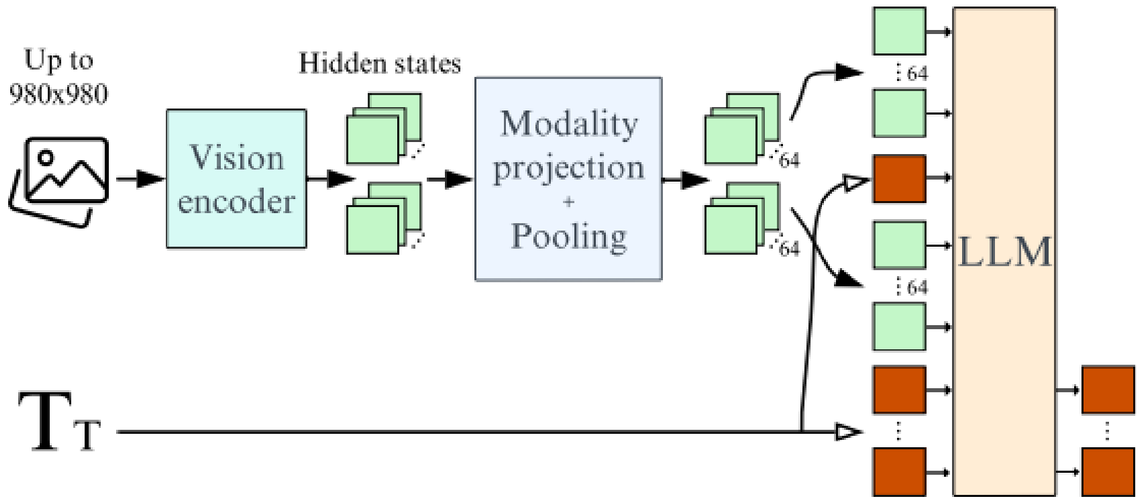
Handling Perturbations

To address the noise caused by adversarial perturbations, features have been extracted in both the frequency and spatial domains.

FURTHER SCOPE

FINE TUNED IDEFICS2 WITH THE TRAINED SIGLIP VISION TRANSFORMER

- The Idefics2 Vision Transformer to be replaced with a SigLip-based Vision Transformer optimized for feature extraction from 32x32 images.
- This SigLip-based model, equipped with a classifier, can be trained for real-fake image and artifact classification, serving as a direct substitute of vision transformer in the Idefics2 framework.



Idefics2 Architecture: The vision encoder to be replaced by trained Siglip Vision Encoder

FURTHER SCOPE

ARTIFACT LOCATION DETECTION

Approach-1: Artifact detection can be done by using models like PaliGemma which can predict bounding boxes and segmentation masks.

Approach-2: Specifically for 32x32 images we can use the patch size of 16, which will give us 4 patches. The cosine similarity between the embedding of one patch and the detected artifacts may give us the location of the artifact.

Thank You