Adobe:
Image Classification and Artifact Detection

Team 86

Overview

Our Aim

Creating a cutting-edge system that detects **AI-generated images** with precision, resists **adversarial attacks**, and explains why the image is generated—setting a new standard for **transparency and reliability**.

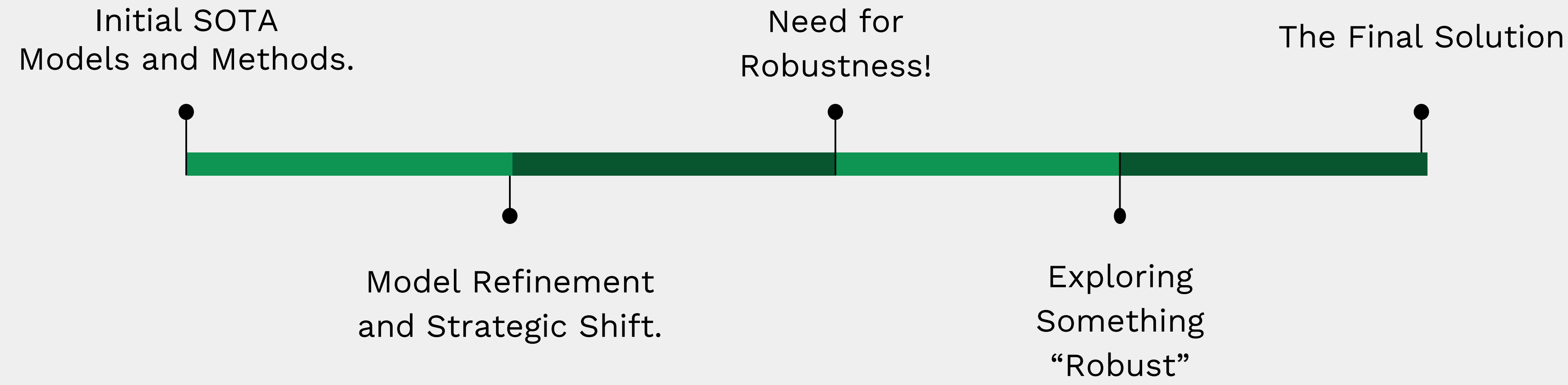
Challenges

- **Visual Ambiguity**: Differentiating real and AI-generated 32×32 images and generating comprehensive explanations based on low-resolution demanded innovative solution.
- **Efficiency Challenge**: Lightweight yet accurate alternatives were essential to replace high-resource State-of-the-art (SOTA) models.
- **Adversarial Robustness**: Robustness against adversarial inputs required synthetic dataset creation.

Preliminary Analysis

Task 01

The Sacred Timeline



Initial SOTA Models and Methods.

CIFAKE Dataset

DIRE

Universal detector for diffusion images via reconstruction error.

57.40% (Accuracy)

AEROBLADE

AI image detector using perceptual loss; struggles with low resolution.

53.81% (Accuracy)

RIGID

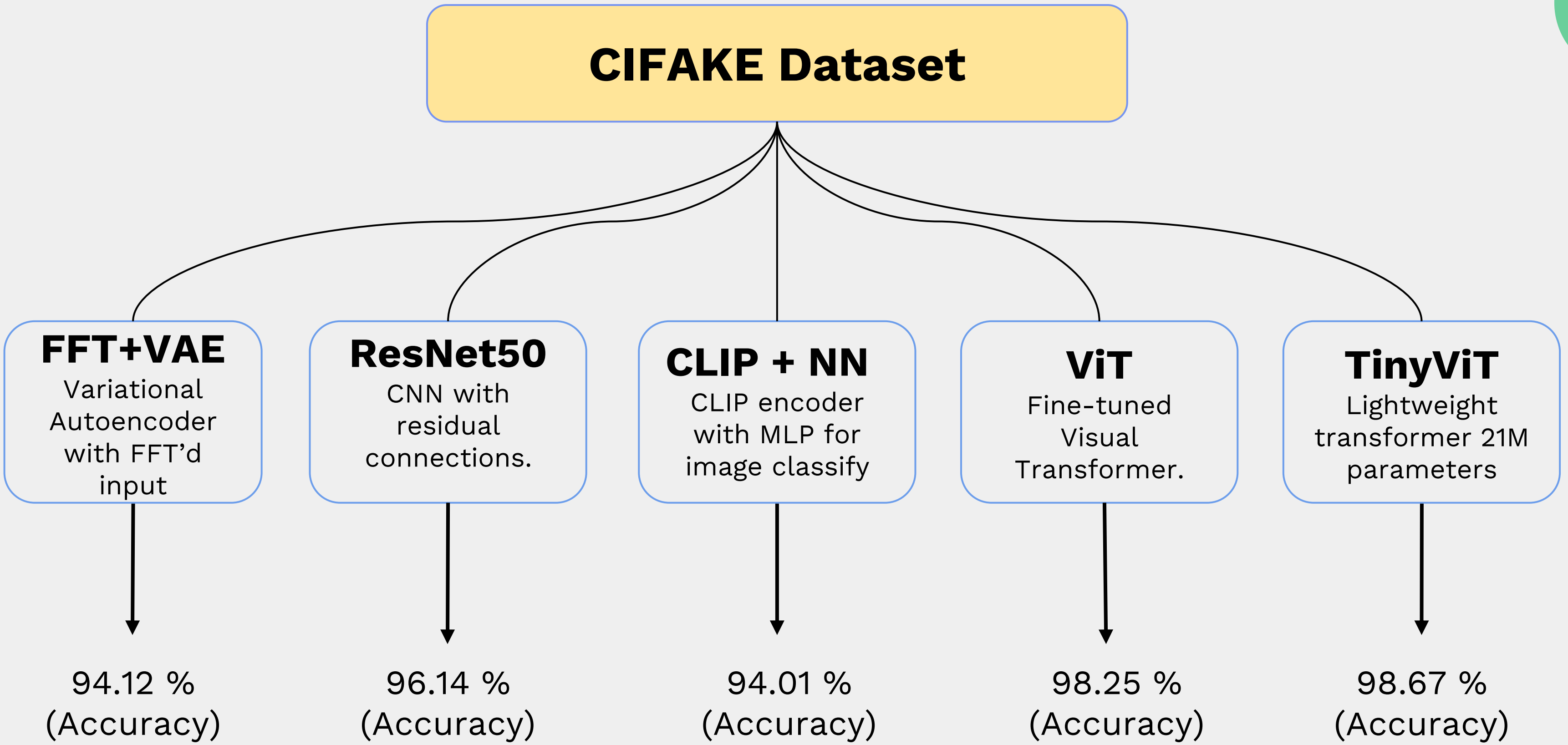
AI image detector exploiting noise sensitivity; suboptimal on CIFAKE.

46.34% (Accuracy)

We intended to explore a different architecture and domain.

Initial SOTA
Models and methods.

Model Refinement and Strategic Shift.


CIFAKE Dataset
FFT+VAE
Variational Autoencoder with FFT’d input
ResNet50
CNN with residual connections.
CLIP + NN
CLIP encoder with MLP for image classify
ViT
Fine-tuned Visual Transformer.
TinyViT
Lightweight transformer 21M parameters
94.12 % (Accuracy)
96.14 % (Accuracy)
94.01 % (Accuracy)
98.25 % (Accuracy)
98.67 % (Accuracy)

Initial SOTA
Models and methods.

Model Refinement
and Strategic Shift.

Need for
Robustness!

We explored 4 attacks

- PGD (Projected Gradient Descent) l_2
- PGD (Projected Gradient Descent) l_∞
- DCT (Discrete Cosine Transform)
- One Pixel Attack

PGD Attack

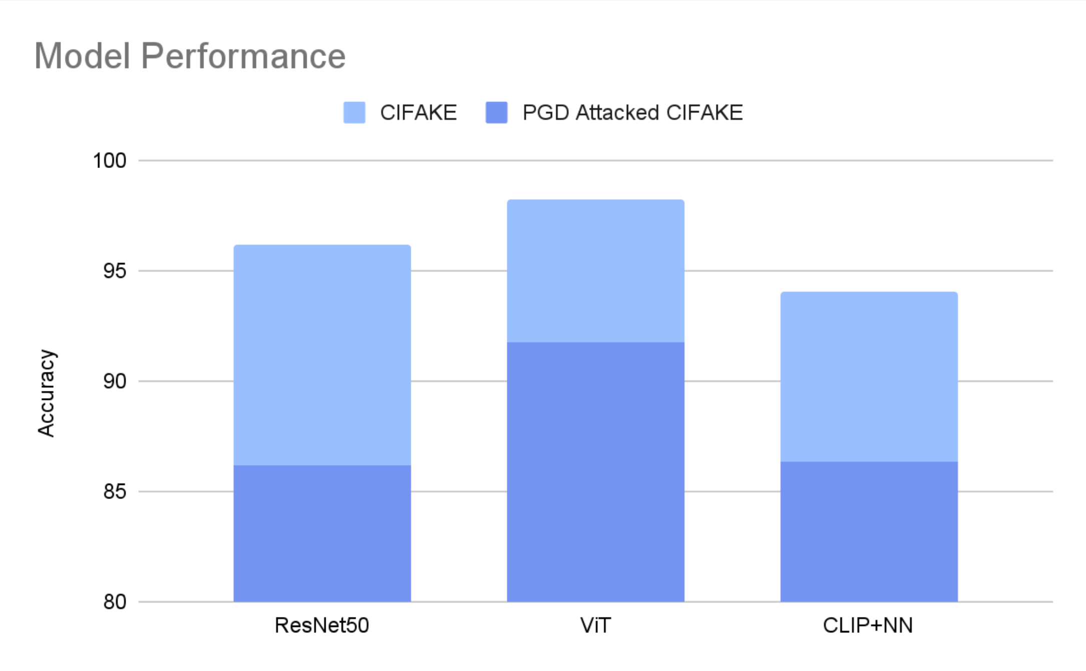
- The PGD attack iteratively adds subtle perturbations to create adversarial examples.
- We implemented ***l_2*-based** (ϵ = 1.0) using ResNet50 and ***l_∞*-based** (ϵ = 16/255) attacks using ViT and ResNet50 on the CIFAKE dataset.

$$\text{Distance} = \sqrt{\sum (x_{\text{new}} - x_{\text{original}})^2}$$ (L2 norm)

$$\text{Distance} = \max\left(\left|x_{\text{new}} - x_{\text{original}}\right|\right)$$ (L_∞ norm)

Real Image | Image scaled to 32*32 | Attacked Image

PGD Attack


Model Performance
CIFAKE
PGD Attacked CIFAKE
100
95
90
85
80
Accuracy
ResNet50
ViT
CLIP+NN

DCT Attack

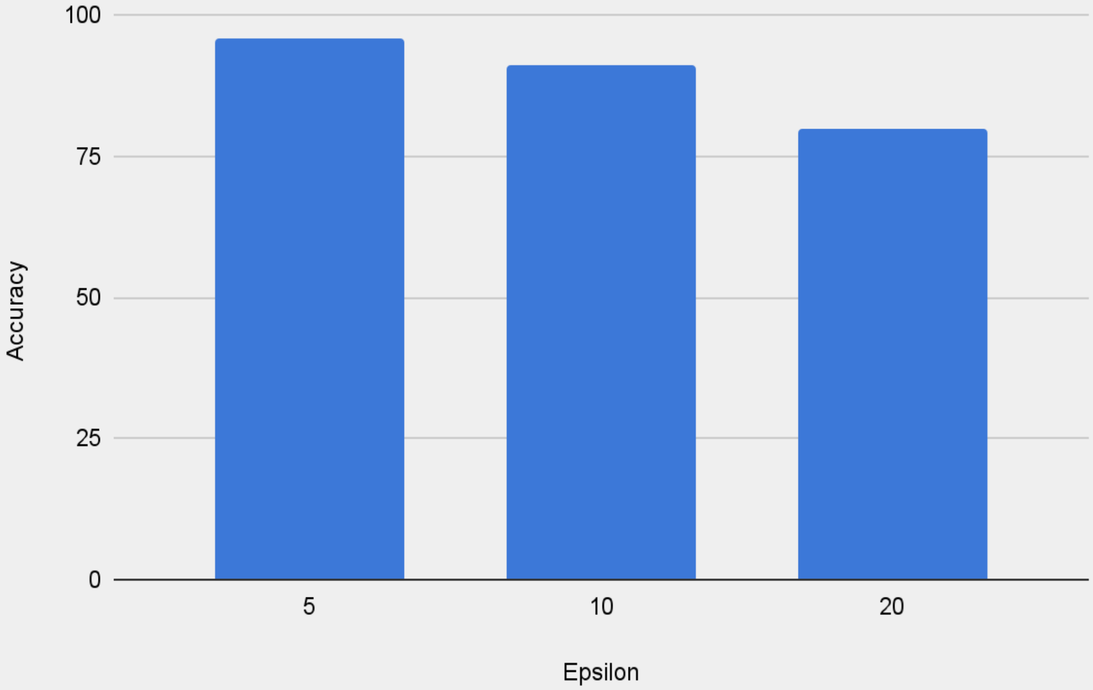
- DCT Attack manipulates the frequency representation of images by introduction of random noise.

Original Image

Epsilon = 5

Epsilon = 15

DCT Attack



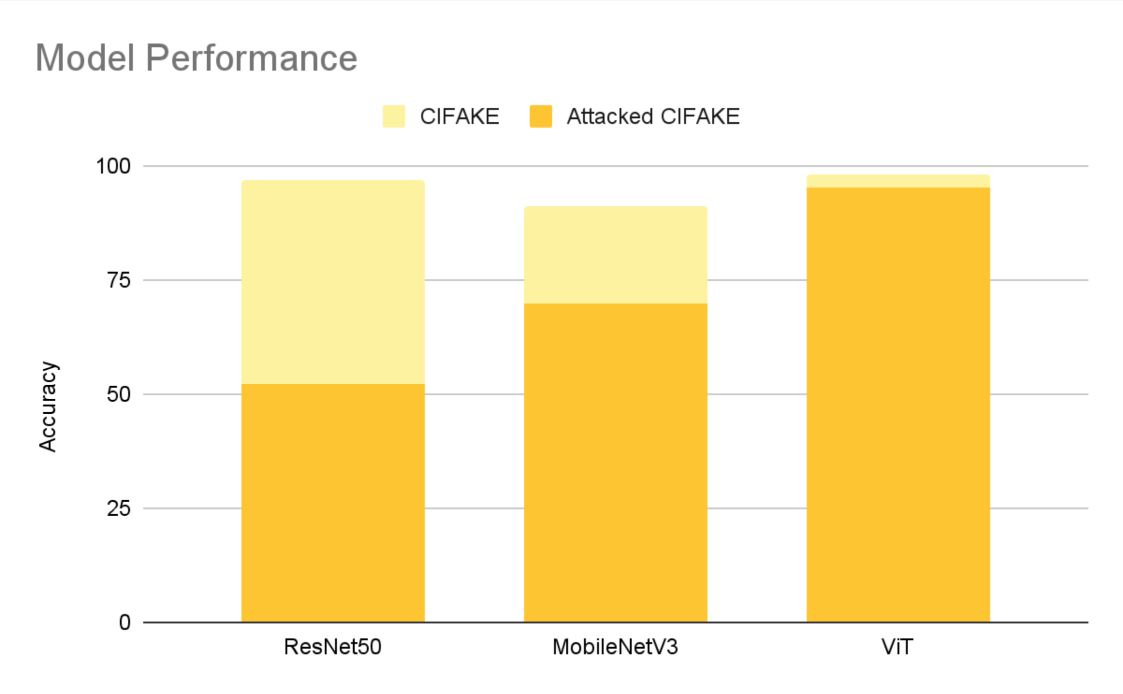
Results on fine-tuned ViT

One Pixel Attack

- Modifies a single pixel to test model robustness and expose vulnerabilities, and aims to cause significant prediction shifts with minimal alterations.
- Reveals the model's sensitivity to pixel-level changes, highlighting weaknesses.

Original Image | 32x32 image | Attacked Image

One Pixel Attack



Perturbed CIFAKE Dataset*

ResNet50
CNN with residual connections.
86.13 % (Accuracy)

CLIP + NN
CLIP encoder with MLP for image classify
86.33 % (Accuracy)

ViT
Fine-tuned Visual Transformer.
91.75 % (Accuracy)

TinyViT
Light-weight transformer 21M parameters
92.87 % (Accuracy)

*dataset consist all type of attacks we explored

Impact of Attack

Models	On CIFAKE	On Adversarial CIFAKE
Resnet50	96.14%	86.13%
CLIP+NN	94.01%	86.33%
ViT	98.25%	91.75%
TinyViT	98.67%	92.87%

Initial SOTA Models and Methods.

Model Refinement and Strategic Shift.

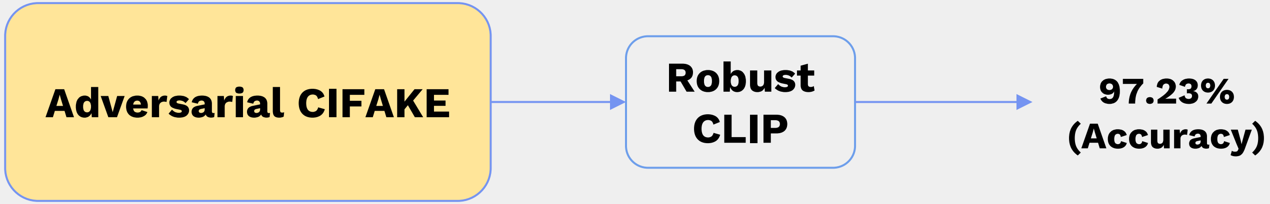
Need for Robustness!

Exploring Something "Robust"

Exploring Something “Robust”

Robust CLIP:

- Fine-tuned variant with resilience to adversarial attacks (**97.23% accuracy on adversarial dataset**).
- Hybrid architecture with CLIP embeddings and lightweight classifier.
- **Struggled with generalization** on CIFAKE and perturbed data, leading to reduced accuracy.



Latency and parameters were far greater than the acceptable limit.

Task 02

Images are Too small!

SOTA models are not able detect artifacts in 32x32 resolution images

Generic Artifact Detection in SOTA Models

- SOTA Vision-Language Models (VLMs) tend to generate **random and overly generic artifacts** when analyzing rescaled images or low-resolution inputs.

Fake Image

GPT-4o Output :

- Blurry and Inconsistent Edges Around Subject
- Repetitive and Unnatural Texture Patterns
- Color Banding and Abrupt Gradient Transitions

Instead of identifying the significant distortion in the shape of the wings, the model provides a more generic description.

VLMs used

SOTA open-source alternatives were evaluated.

For example:

- PaliGemma-3B
- Qwen2-VL-7B-Instruct
- Llama-3.2-11B-Vision

Model Outputs

Llama-3.2 11B Vision outputs:

- Unrealistic lighting, highlights are too sharp and lack natural diffusion.
- **Fur lacks realistic texture** and individual hair detail, appears overly smooth.
- Background trees are **blurry and lack depth of field consistency.**
- Colors are overly saturated, especially the deer's coat.

Upscaling for Enhanced Artifact Detection

- To improve the Vision-Language Model (VLM)'s ability to detect artifacts, fake images were upscaled using various advanced upscalers.

- The upscaling process **should enhance** details, make artifacts more pronounced and easier for the VLM to identify and analyze.

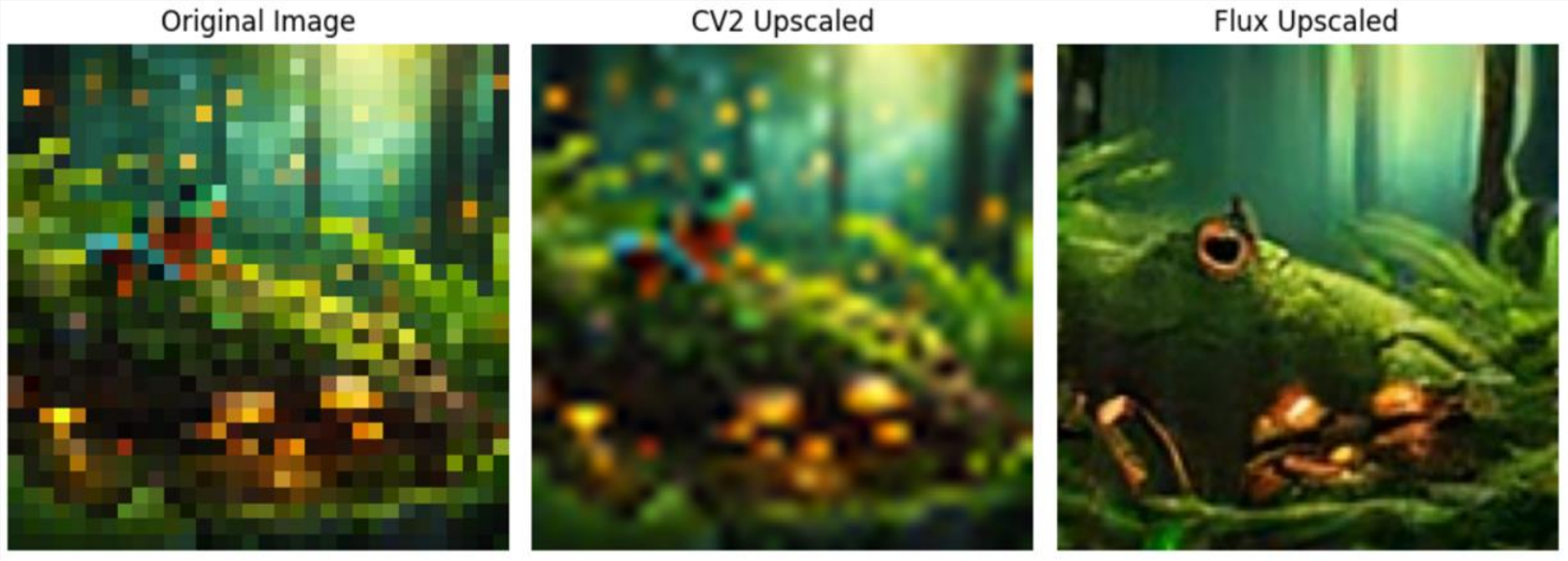
- The following upscaling techniques were used:
    - **OpenCV upscaling**
    - **Flux-upscaler**

Upscaling Introduces

Unwanted Artifacts !

The upscaling process itself introduces new artifacts that may not be present in the original images.


Original Image
CV2 Upscaled
Flux Upscaled


Original Image
CV2 Upscaled
Flux Upscaled

Proposed Solution

Dataset

Dataset

CIFAKE ++
(Task 01)

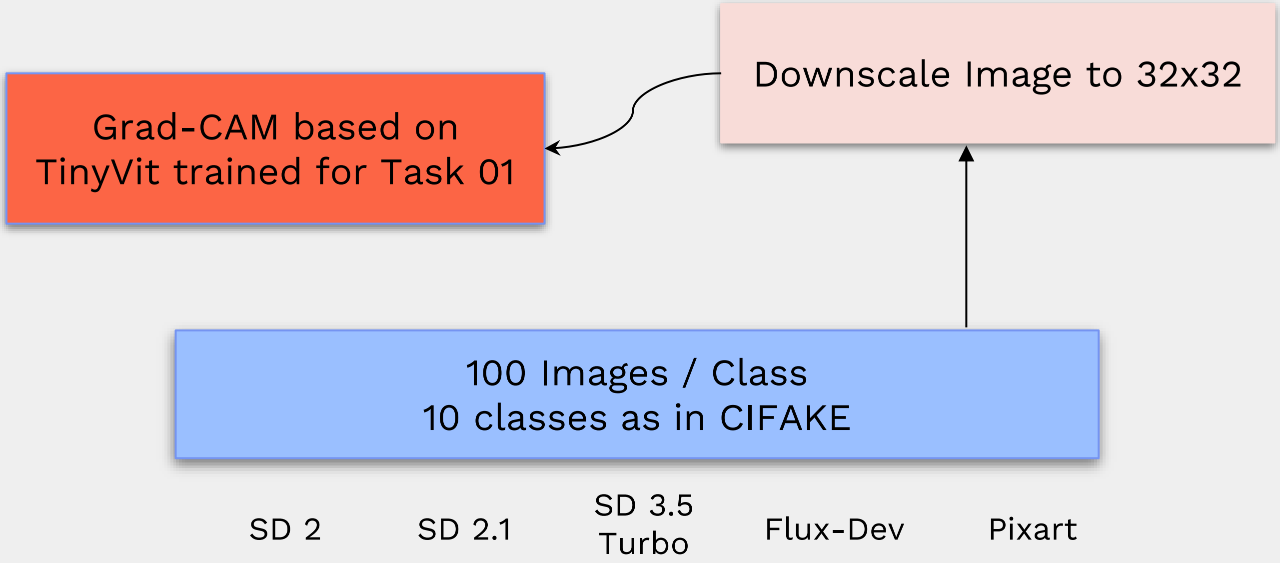
Image-Explanation Pair
(Task 02)

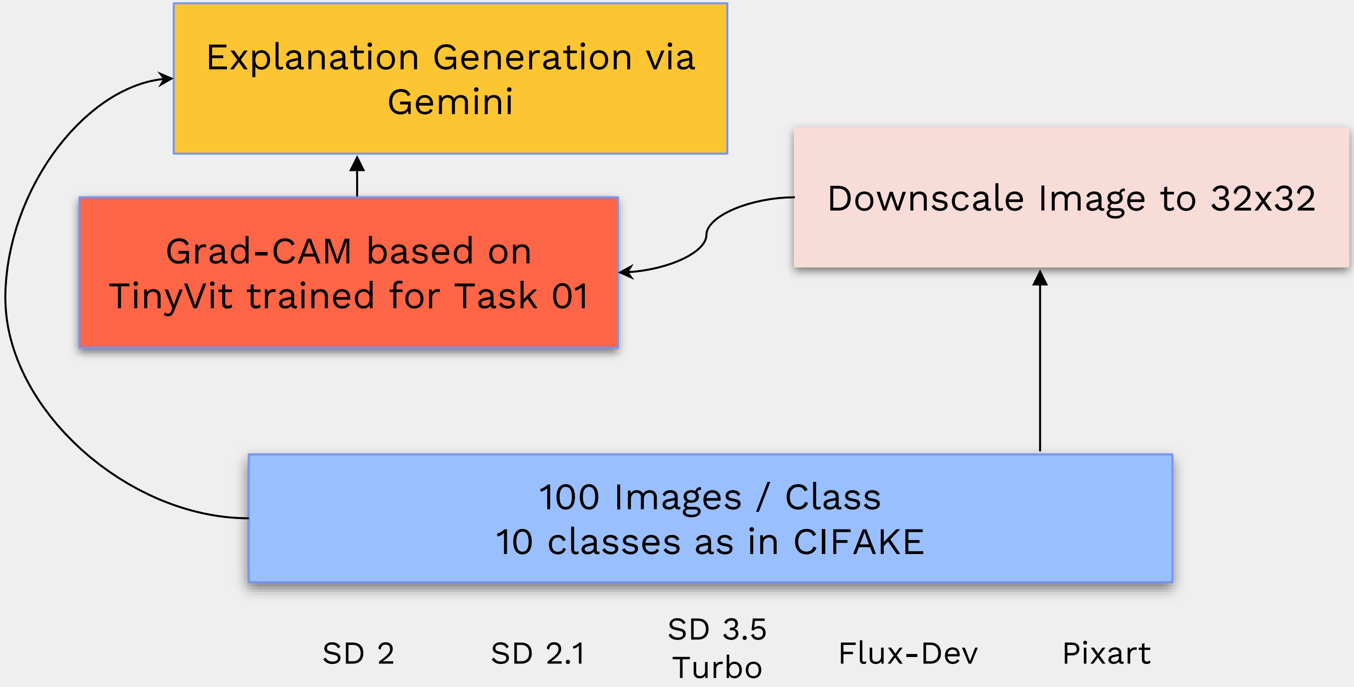
CIFAKE++

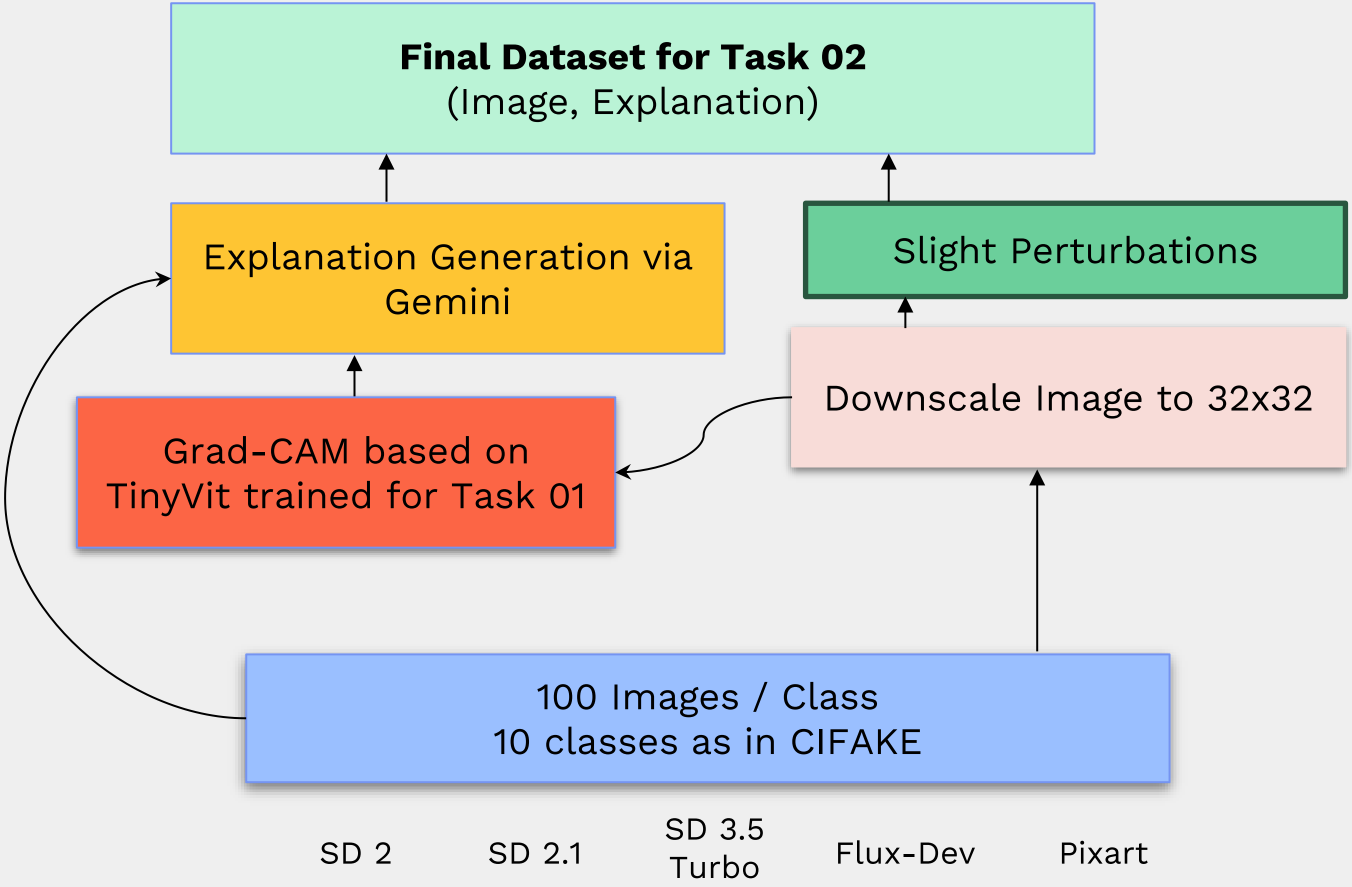
- Original CIFAKE
- Adversarial CIFAKE
  - PGD-l_2
  - PGD-l_∞
  - DCT
  - One-Pixel-Attack
- Generated Images
  - SD 2
  - SD 2.1
  - SD 3.5 Turbo
  - Flux-dev
  - PixArt

100 Images / Class
10 classes as in CIFAKE

SD 2 | SD 2.1 | SD 3.5 Turbo | Flux-Dev | Pixart


Downscale Image to 32x32
Grad-CAM based on
TinyVit trained for Task 01
100 Images / Class
10 classes as in CIFAKE
SD 2
SD 2.1
SD 3.5
Turbo
Flux-Dev
Pixart


Explanation Generation via Gemini
Grad-CAM based on TinyVit trained for Task 01
Downscale Image to 32x32
100 Images / Class
10 classes as in CIFAKE
SD 2
SD 2.1
SD 3.5 Turbo
Flux-Dev
Pixart


Final Dataset for Task 02
(Image, Explanation)
Explanation Generation via Gemini
Slight Perturbations
Grad-CAM based on TinyVit trained for Task 01
Downscale Image to 32x32
100 Images / Class
10 classes as in CIFAKE
SD 2
SD 2.1
SD 3.5 Turbo
Flux-Dev
Pixart

Prompt Used For Gemini

Prompt designed for concise and accurate explanations for each fake image:

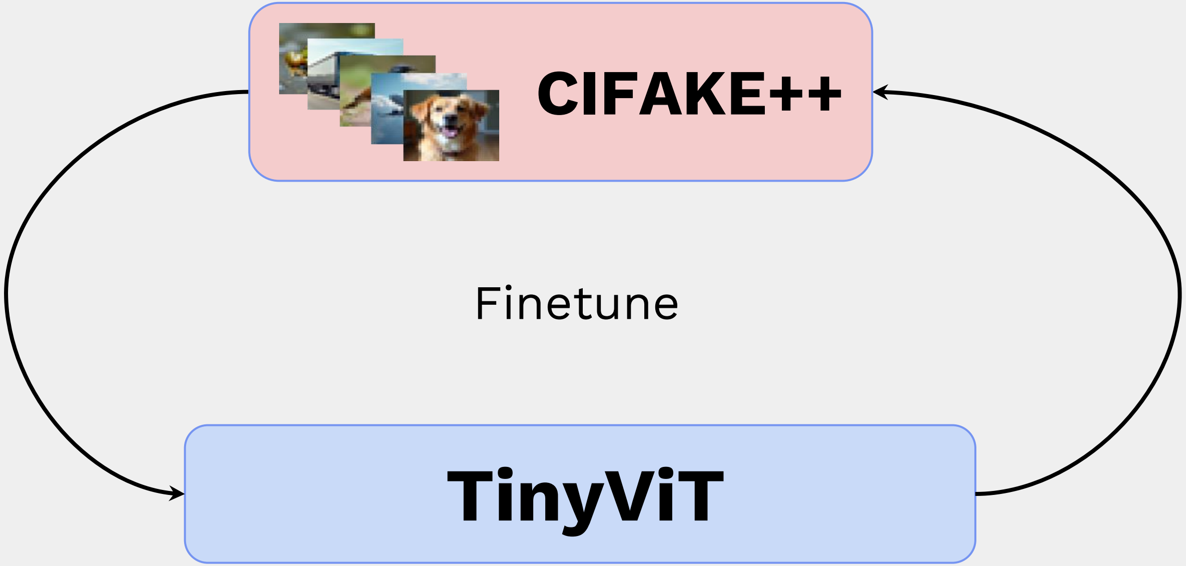
Analyze the provided image, and its corresponding Grad-CAM output which has been resized to 32x32. Focus primarily on the original image to identify and explain distinguishing artifacts that indicate it is fake. Use the Grad-CAM output for reference only when necessary. Provide clear, concise explanations (maximum 50 words each) using the specified artifacts below. Include positional references like 'top left' or 'bottom right' when relevant. DO NOT include any other sentences or artifacts in your response. Select only 6-7 relevant artifacts.

<LIST OF ARTIFACTS>....</LIST OF ARTIFACTS>

Methodology
(Task 01)

Robust TinyViT

TinyViT **fine-tuned** on the **CIFAKE ++** dataset to create **Robust TinyVit**.



Robust TinyViT : Results

Dataset	Accuracy
CIFAKE	98.67%
CIFAKE++	98.61%

Latency: **35 milli-seconds per image** (batch size 32) and **75 milli-seconds for single-image inference** on server-grade **CPU.**

Methodology
(Task 2)

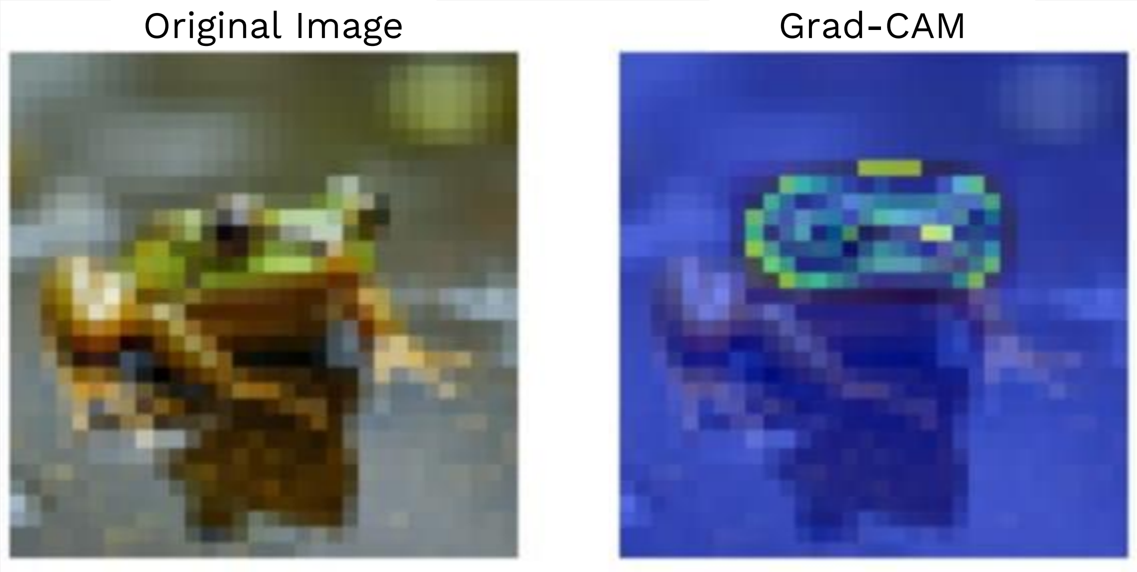
Fine-tuning Qwen2-VL (7B)

Qwen2-VL was fined tuned on the **Image-Explanation pairs** obtained as described earlier using **Low-Rank-Adapters (LoRA)**.

- **Why Qwen2-VL?**
  - Strong visual-understanding capabilities
  - Ability to handle multiple images simultaneously
- **Input?**
  - Actual 32x32 image
  - Grad-CAM as obtained from fine-tuned TinyViT

Why Grad-CAM ?

Grad-CAM helps the model to be more positionally aware of where potential artifacts might be present.

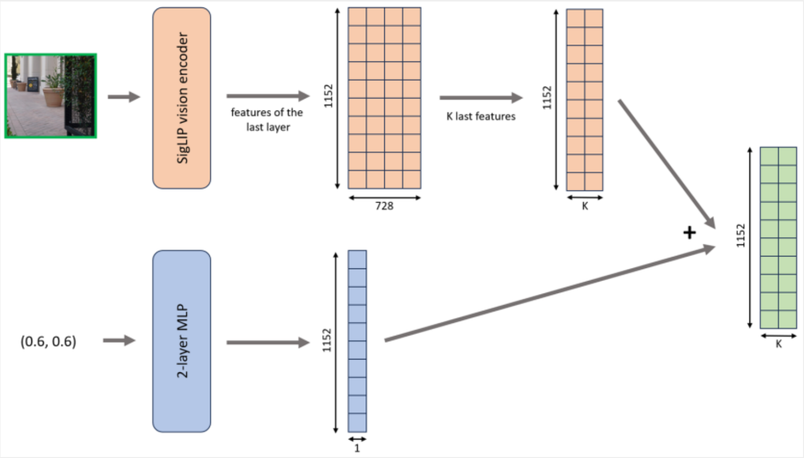


Fine-tuning MC-LLaVA (3B)

Pretrained MC-LLaVA was **fine-tuned** on the **Image-Explanation pairs** obtained as described earlier using **Low-Rank-Adapters (LoRA)**.

- **Why MC-LLaVA?**
  - The unique training objective and style of MC-LLaVA, based on dividing a high resolution images into multiple low resolution sections in order to help the model **"Zoom"** into specific parts of the image.

Enabling VLMs to "Zoom"



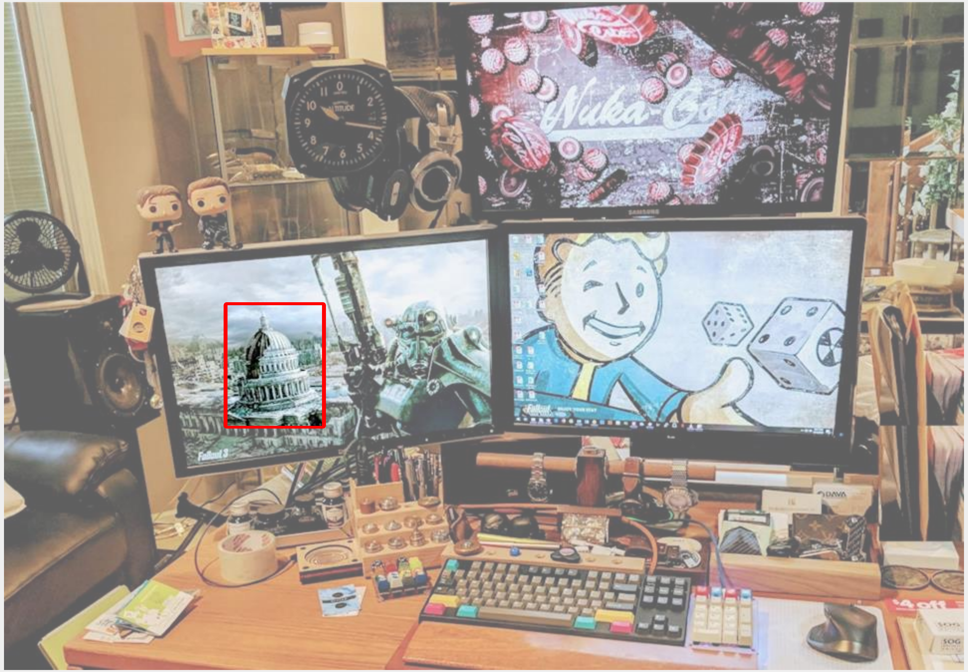
https://huggingface.co/blog/visheratin/vlm-resolution-curse

Our Hypothesis

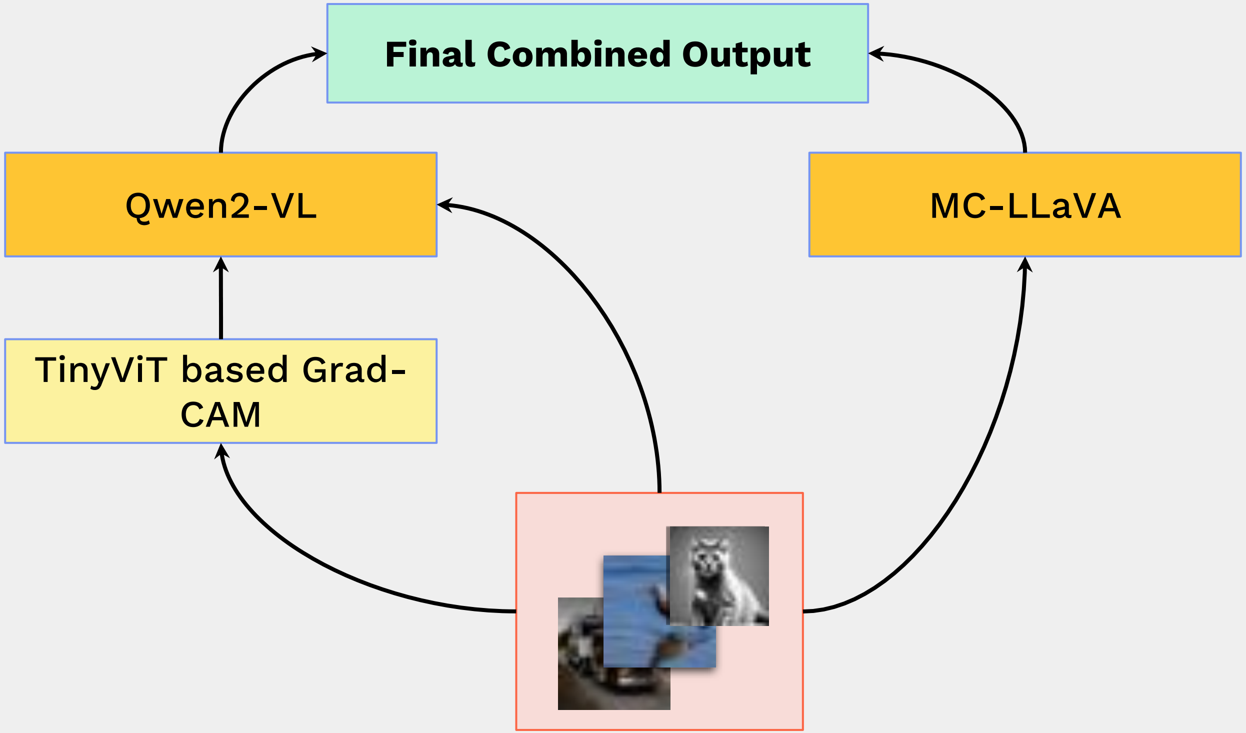
The 32x32 images can be considered as a **part of a larger image** and our task is to analyse this small section.

Our Hypothesis

Hypothesis: The 32x32 images can be considered as **a part of a larger image** and our task is to analyse this small section.



The “Multi-Model” Setup



An example for Task 2

Original Image

Downscaled Image

Unrealistic eye reflections: Symmetrical, overly bright reflections in the eyes lack natural irregularity.

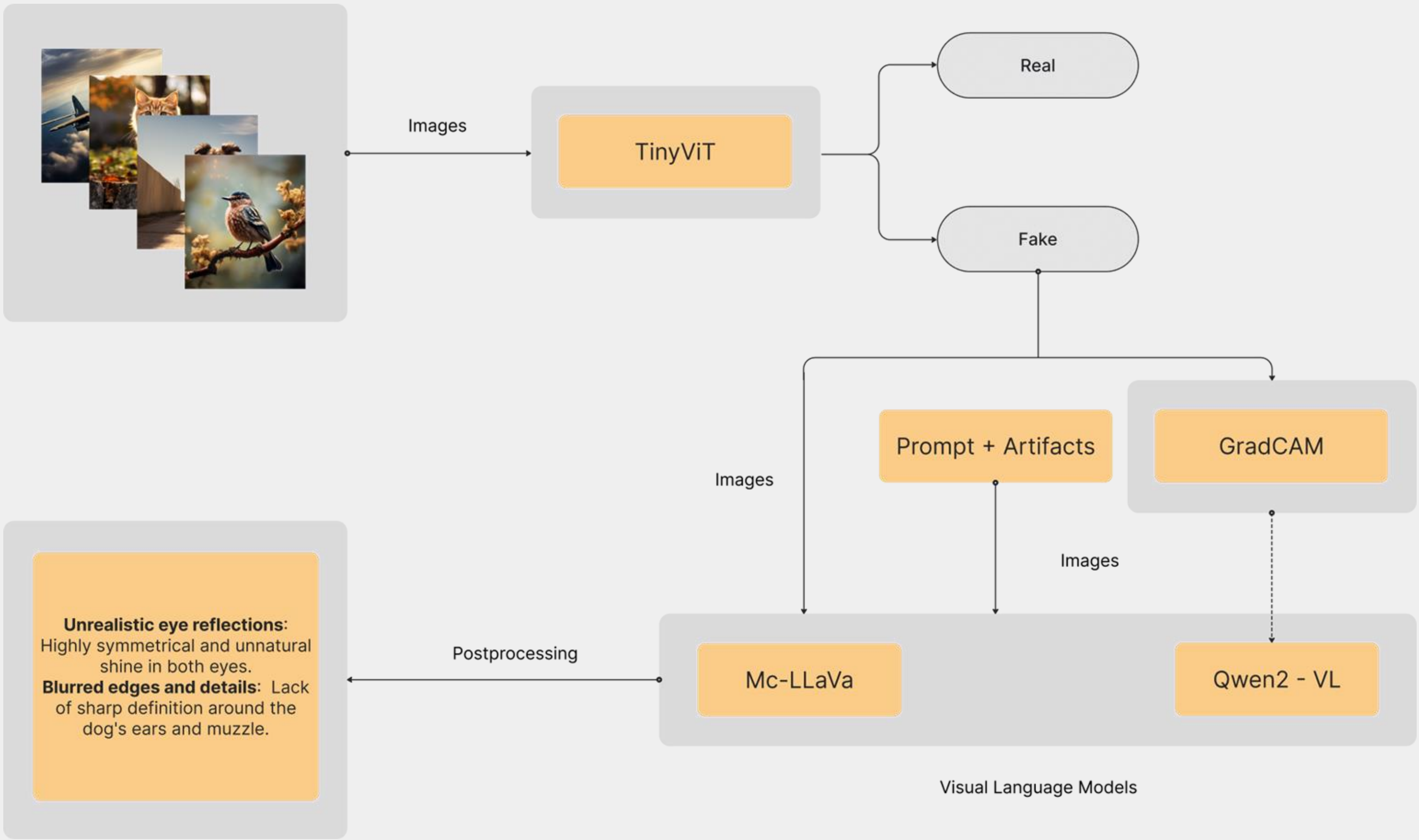
Over smoothening of Natural Texture: The frog's skin appears unnaturally smooth and lacks the natural bumps and irregularities of real skin.

Inconsistent lighting: The lighting on the frog is uneven and lacks natural shadows, suggesting artificial lighting.

Color coherence break: The frog's colors are overly saturated and vibrant, lacking the subtle variations of real amphibian skin tones.

Fake depth of field: The frog's details are sharp, but the background is blurry and lacks detail.

The Final Flowchart


Images
TinyViT
Real
Fake
Images
Prompt + Artifacts
GradCAM
Images
Unrealistic eye reflections: Highly symmetrical and unnatural shine in both eyes.
Blurred edges and details: Lack of sharp definition around the dog's ears and muzzle.
Postprocessing
Mc-LLaVa
Qwen2 - VL
Visual Language Models

Summary!

- **Building the CIFAKE++ Dataset :** CIFAKE + Adversarially Perturbed Images + Synthetic Images generated using Diffusion Models.
- **Robust TinyViT Excelled in Task 01:** AI-Generated Image Classification, delivering strong performance on both clean and adversarial datasets with reduced latency and computational overhead.
- **Building Dataset for Task 02:** Generation of explanations via Gemini and Grad-CAM based on Robust TinyViT, trained in Task 01. Images are downscaled to 32x32, with 100 images per class, sourced from various diffusion models.
- **"Multi-Model" Pipeline:** The combination of MC-LLaVA and Qwen2-VL paired with Grad-CAM yielded the desired explanation of the artifacts present in the image.

Thanks!

References

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., & Li, H. (2023). Dire for diffusion-generated image detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22445-22455).	DIRE for Diffusion-Generated Image Detection
Ricker, J., Lukovnikov, D., & Fischer, A. (2024). AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9130-9140).	AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error
He, Z., Chen, P. Y., & Ho, T. Y. (2024). RIGID: A Training-free and Model-Agnostic Framework for Robust AI-Generated Image Detection. *arXiv preprint arXiv:2405.20112*.	RIGID: A Training-free and Model-Agnostic Framework for Robust AI-Generated Image Detection
Madry, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.	Towards Deep Learning Models Resistant to Adversarial Attacks

References

Jia, S., Ma, C., Yao, T., Yin, B., Ding, S., & Yang, X. (2022). Exploring frequency adversarial attacks for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4103-4112).	Exploring Frequency Adversarial Attacks for Face Forgery Detection
Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828-841.	One pixel attack for fooling deep neural networks
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.	Learning Transferable Visual Models From Natural Language Supervision
Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

References

Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., & Yuan, L. (2022, October). Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision* (pp. 68-85). Cham: Springer Nature Switzerland.	TinyViT: Fast Pretraining Distillation for Small Vision Transformers
Schlarmann, C., Singh, N. D., Croce, F., & Hein, M. (2024). Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*.	Robust CLIP: Unsupervised Adversarial Fine-Tuning of Vision Embeddings for Robust Large Vision-Language Models
Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., ... & Lin, J. (2024). Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.	Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution
An, R., Yang, S., Lu, M., Zeng, K., Luo, Y., Chen, Y., ... & Zhang, W. (2024). MC-LLaVA: Multi-Concept Personalized Vision-Language Model. *arXiv preprint arXiv:2411.11706*.	MC-LLaVA: Multi-Concept Personalized Vision-Language Model