University of the
Philippines Los Baños

AI-NO SWIPING

Adversarial Perturbation Tool to Protect Digital Artworks from AI Misuse

SPECIAL PROBLEM
CMSC 190

PRESENTED BY:
John Lawrence F. Quiñones

DALL·E

Midjourney

Text-to-Image Generative Models

Relies heavily on large datasets. Datasets comprise of millions of images scraped from the internet (e.g. LAION dataset).

Robert Kneschke vs LAION

Background of the Study

“Ghiblifaction”

Background of the Study

Public Availability ≠ Unrestricted Use

Custom fine-tuning of models

INDIVIDUALS CAN TAKE ARTWORKS ONLINE AND USE THEM TO FINE-TUNE PERSONALIZED MODELS.

FINETUNED MODELS CAN RECREATE ARTIST'S STYLE.

What can Artists do?

1

Opting out

Offered by AI companies

2

Image Protection Tools

Adversarial tools

1

Opting out

Manually register objections for each included work.

Tedious and impractical.

What can Artists do?

1

Opting out

Offered by AI companies

2

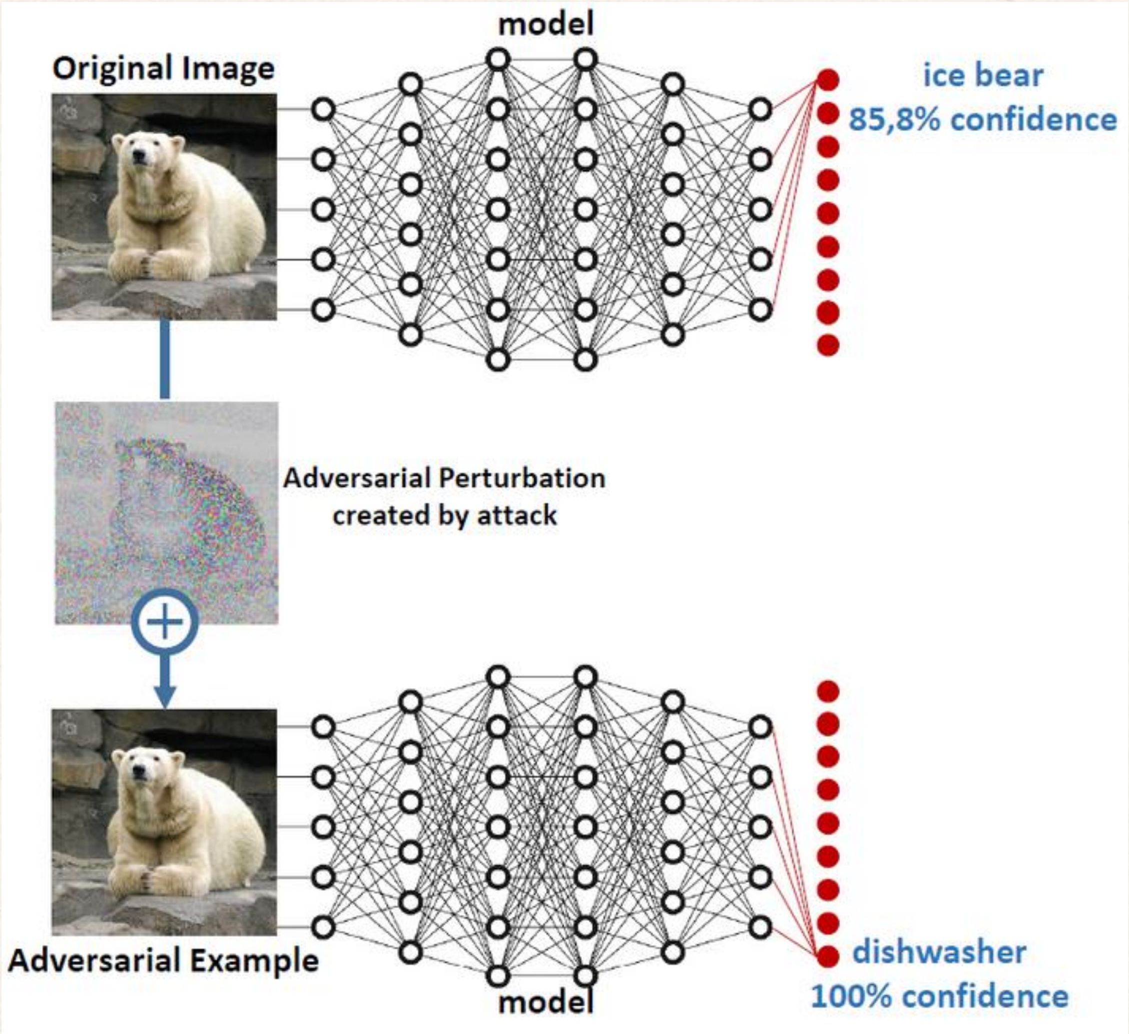
Image Protection Tools

Adversarial tools

2

Image Protection Tools

Leverages *Adversarial Perturbations*



"subtle modifications introduced to input data that can significantly mislead machine learning models into making incorrect predictions"

2

Image Protection Tools

The Problem?

Image protection tools exist, but they are inaccessible to most artists.

2

Image Protection Tools

High hardware demands.

Tools	Overview
Glaze/Nightshade	Desktop App, 5-6 GiB VRAM
WebGlaze	Web-based, requires account creation (emailing the creators, submitting art portfolio for proof of artistry)
Mist	NVIDIA RTX 3090 GPU
Anti-Dreambooth	NVIDIA A100 GPU
Dormant	Intel Xeon Gold 5218R CPU, 4 NVIDIA 1800 (80GB) GPU’s
CAAT	NVIDIA RTX 3090 GPU
DIAGNOSIS	6 Quadro RTX 6000 GPU’s

2

Image Protection Tools

Only few are available as stand-alone applications, most are research code.

Objectives

This study aims to:

Objective #1

Create a memory efficient adversarial perturbation algorithm.

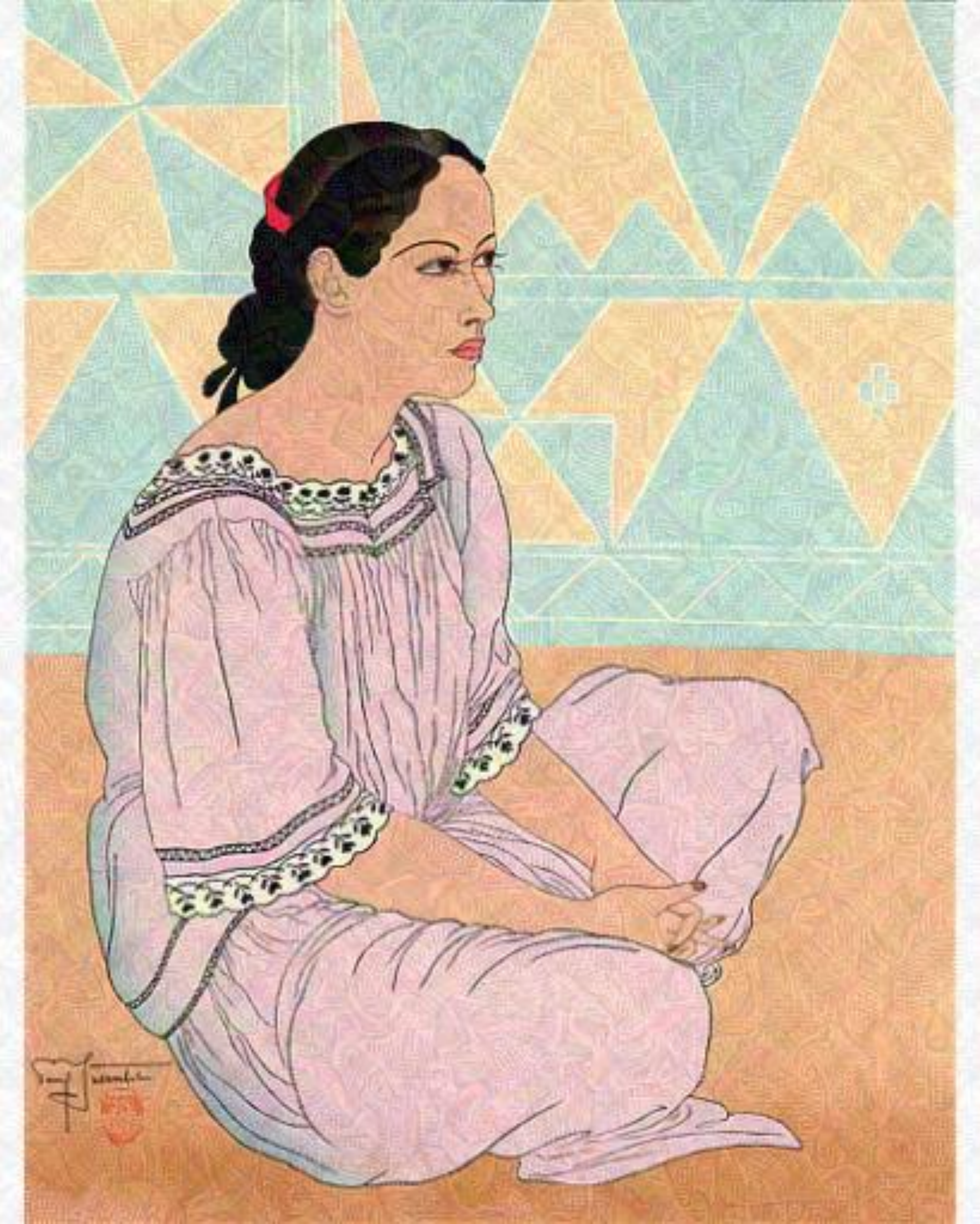
Objective #2

Integrate the perturbation algorithm to a custom desktop application.

Objective #3

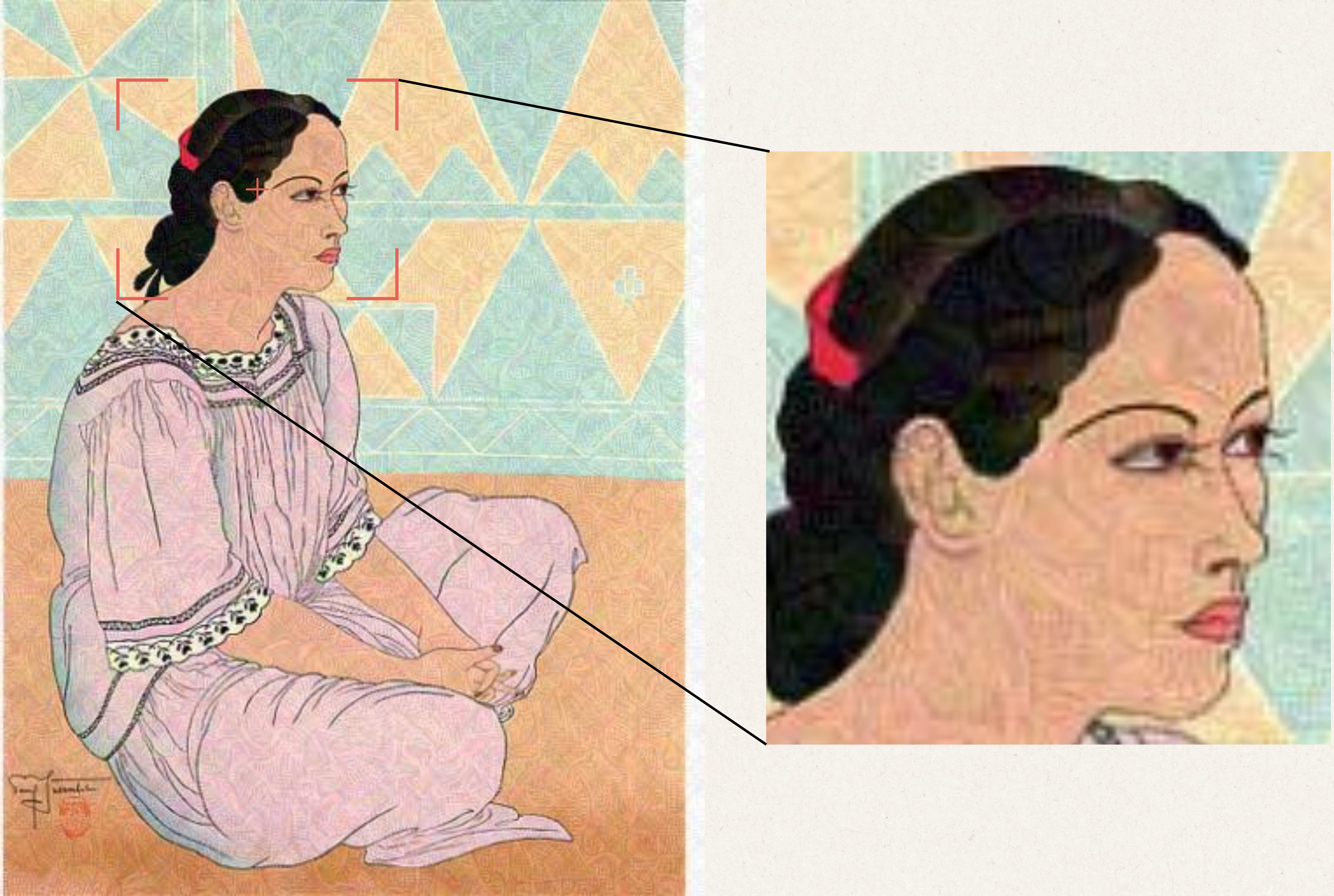
Test the effectiveness of developed tool against a locally trained diffusion model.

AINS' Perturbation Algorithm



Perturbed Image Using AINS

Painting by Paul Jacoulet



Techniques Used

Projected Gradient Descent (PGD)

Based on Anti-Dreambooth’s Alternating Surrogate Perturbation Learning (ASPL) approach

Techniques Used

Projected Gradient Descent (PGD)

Based on Anti-Dreambooth's Alternating Surrogate Perturbation Learning (ASPL) approach

Image Tiling

Using Dask library

Techniques Used

Projected Gradient Descent (PGD)

Based on Anti-Dreambooth's Alternating Surrogate Perturbation Learning (ASPL) approach

Image Tiling

Using Dask library

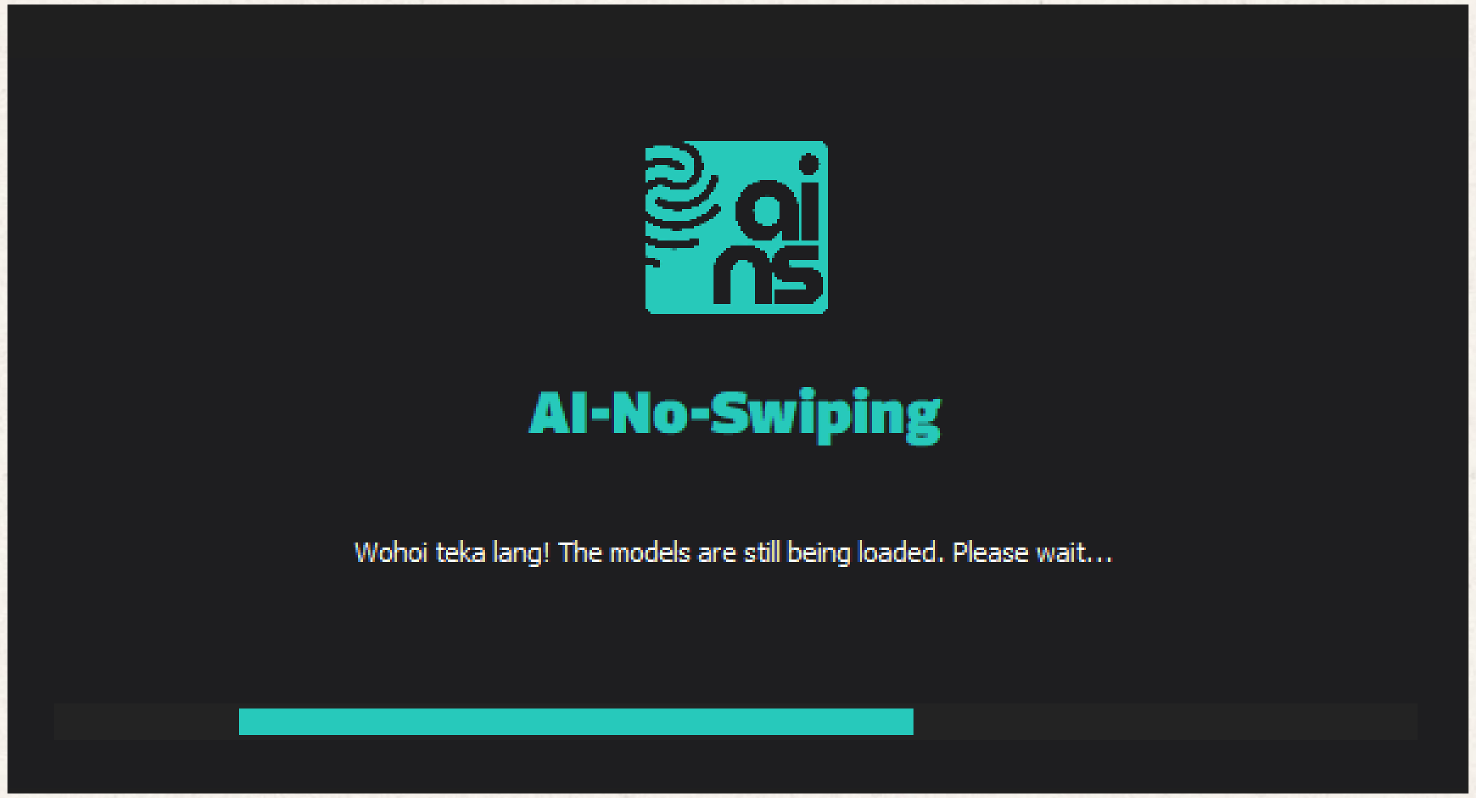
Half Precision Model Loading

UNet, VAE, Text Encoder

AINS Desktop App

AINS App

Developed using PyQt5 GUI framework.



Splash Screen

Loading models.

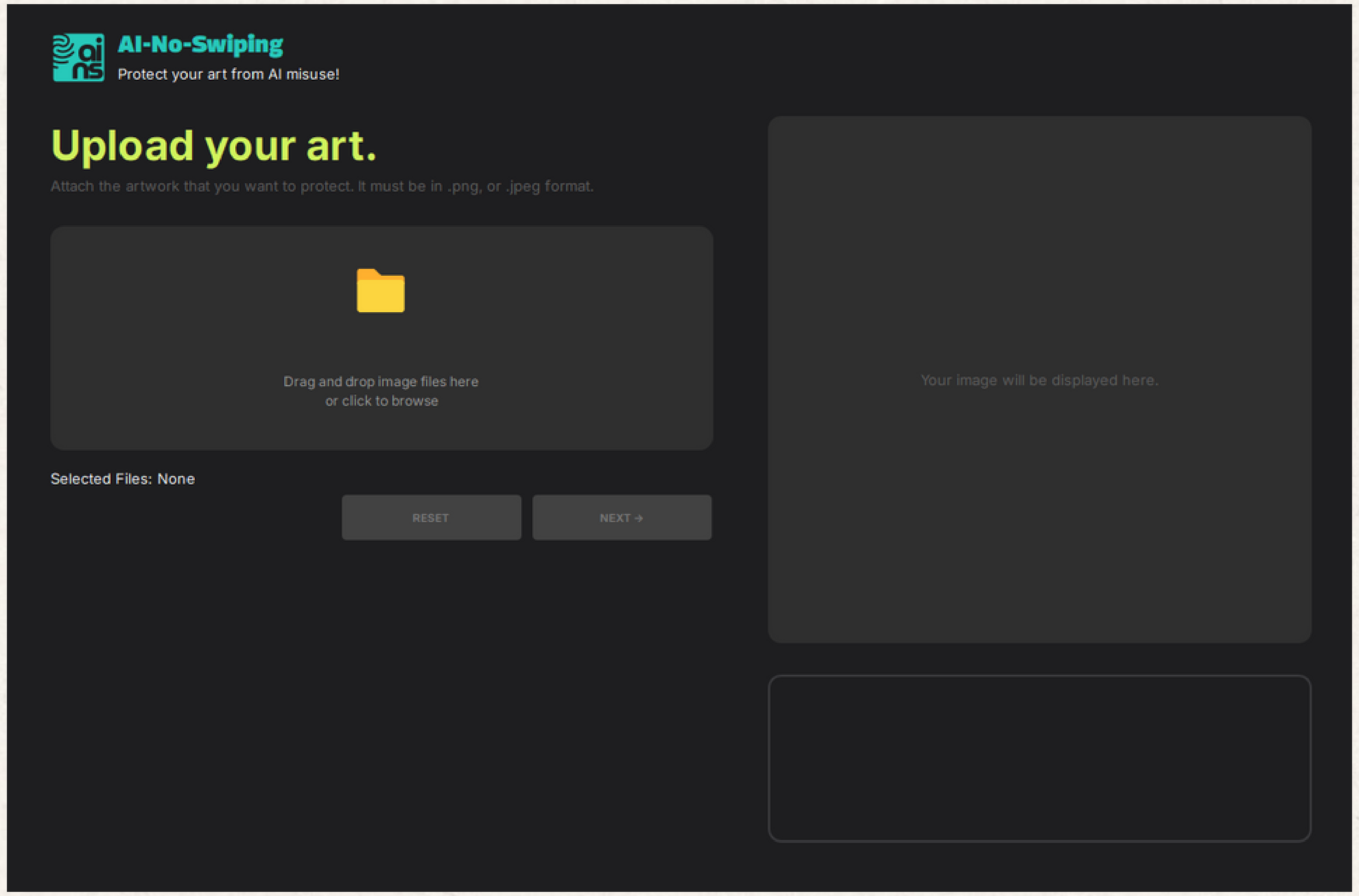
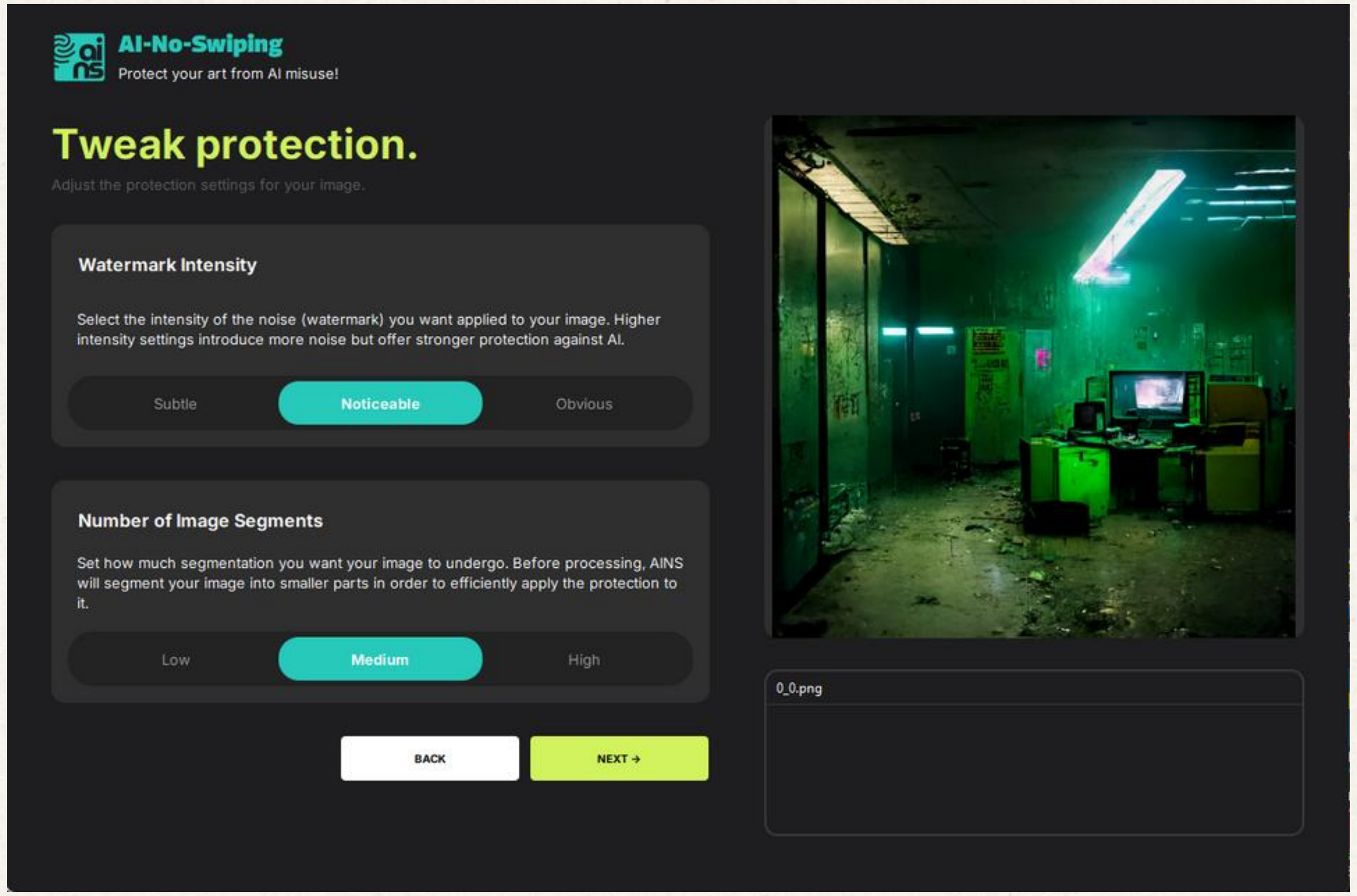


Image Selection Screen

Select one or more images.

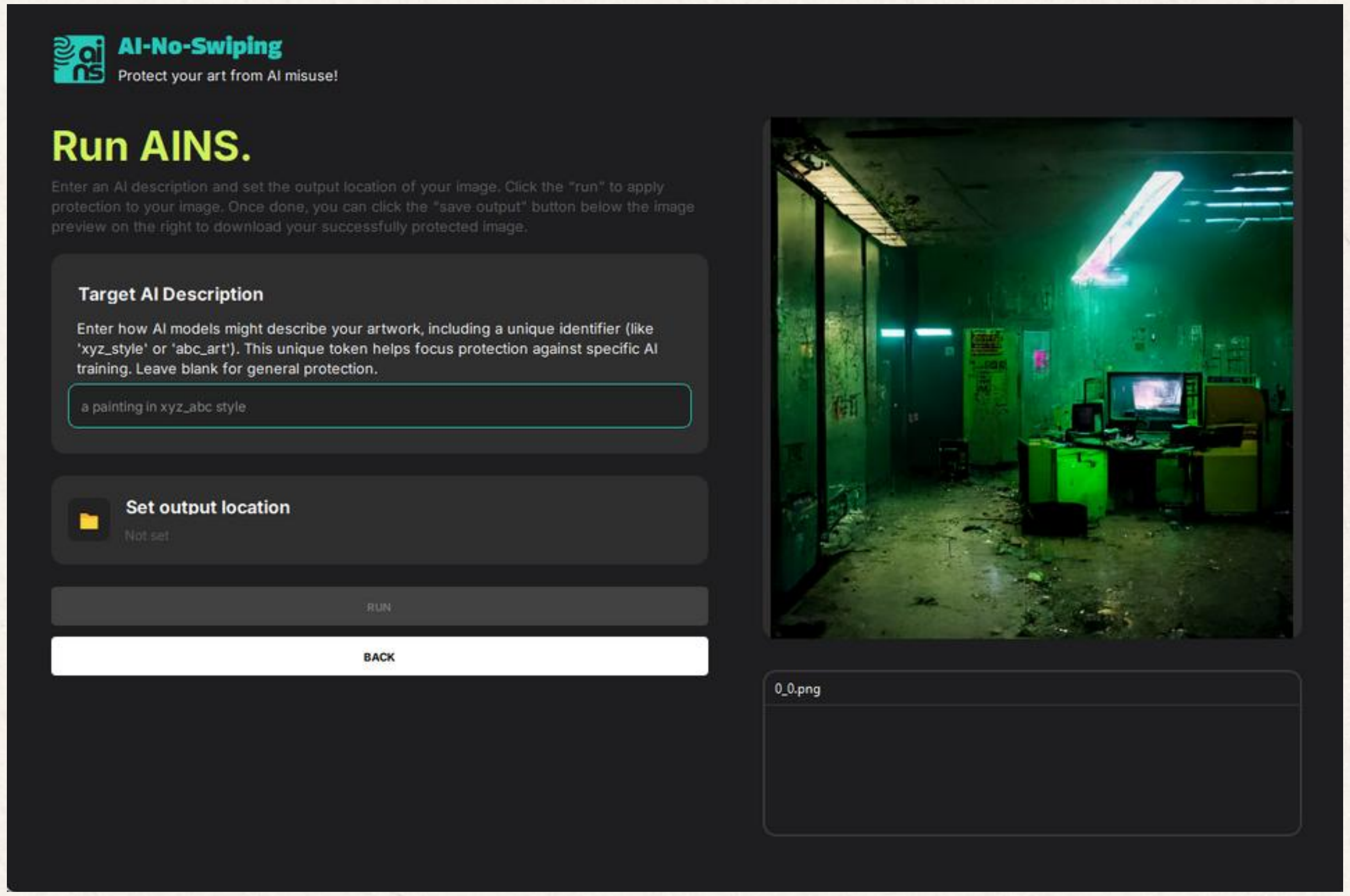
AINS App

Developed using PyQt5 GUI framework.



Perturbation Configuration Screen

Set Watermark & Tiling intensity.



Instance Prompt and Output configuration Screen

Optionally Instance Prompt & Output location. Run perturbation.

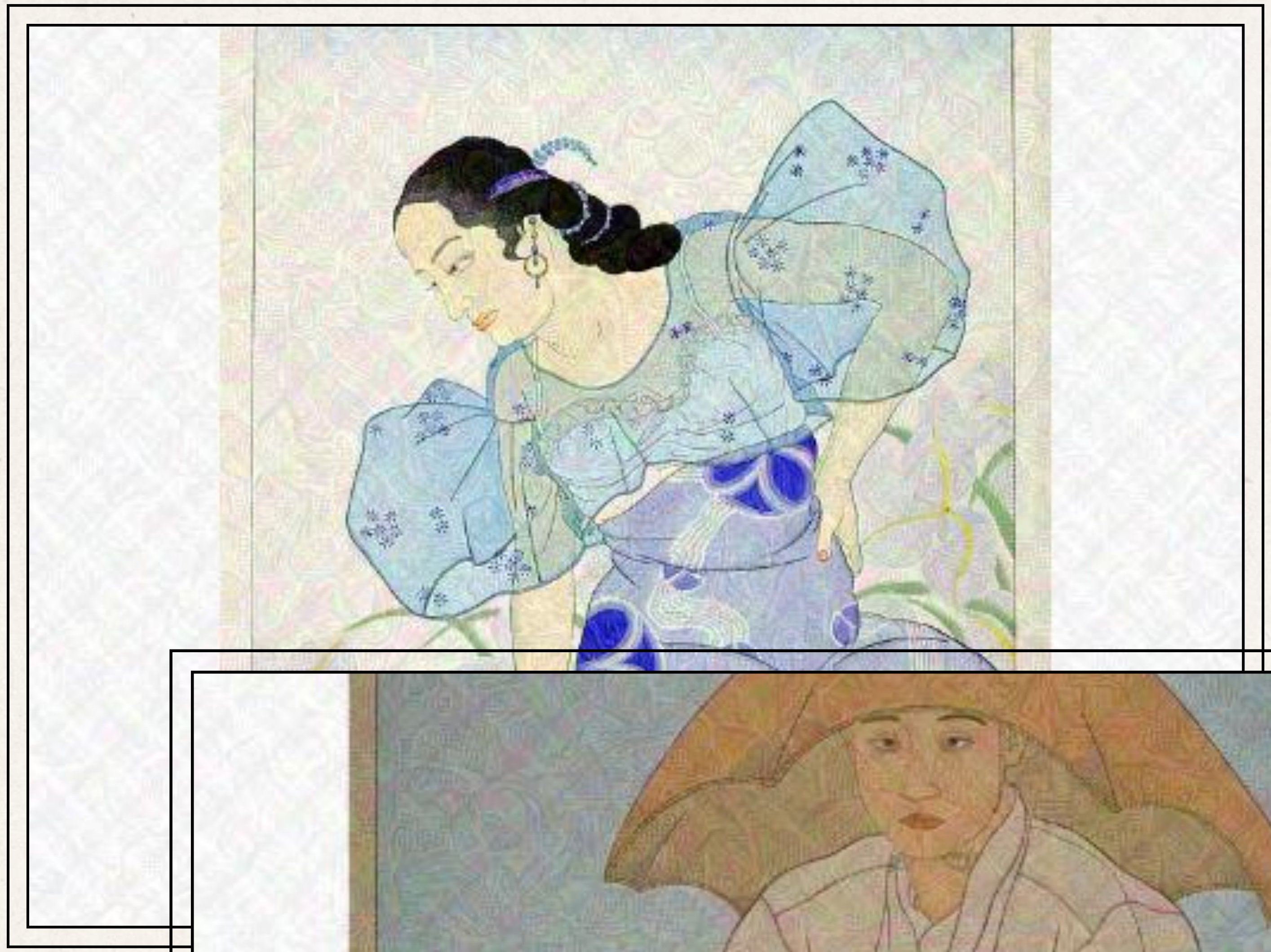
Effectiveness and Efficiency Evaluation

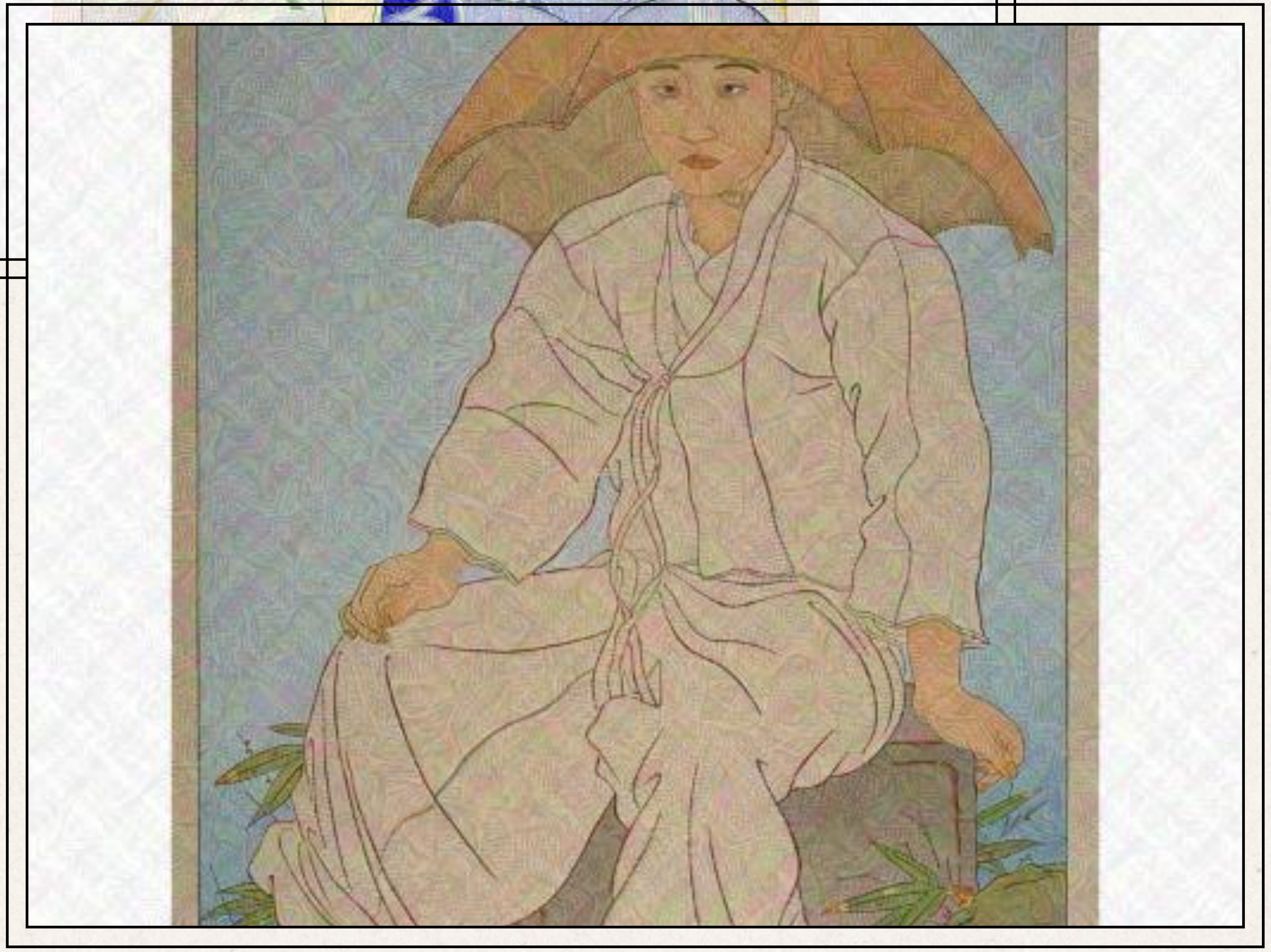
Adversarial Attack + Resource Usage

Experiment Setup

*Representative only. Not full dataset used.

30 Images of Paul Jacoulet's paintings.





Experiment Setup

Perturbation Settings Applied:

- ✦ **Watermark Intensity:** Obvious
- ✦ **Tiling Intensity:** High

Resource Usage:

- ✦ GPU VRAM: 3.5 GiB (idle), 3.8 GiB (active perturbation)
- ✦ RAM: 1.634 GiB (idle), 1.645 (active perturbation)

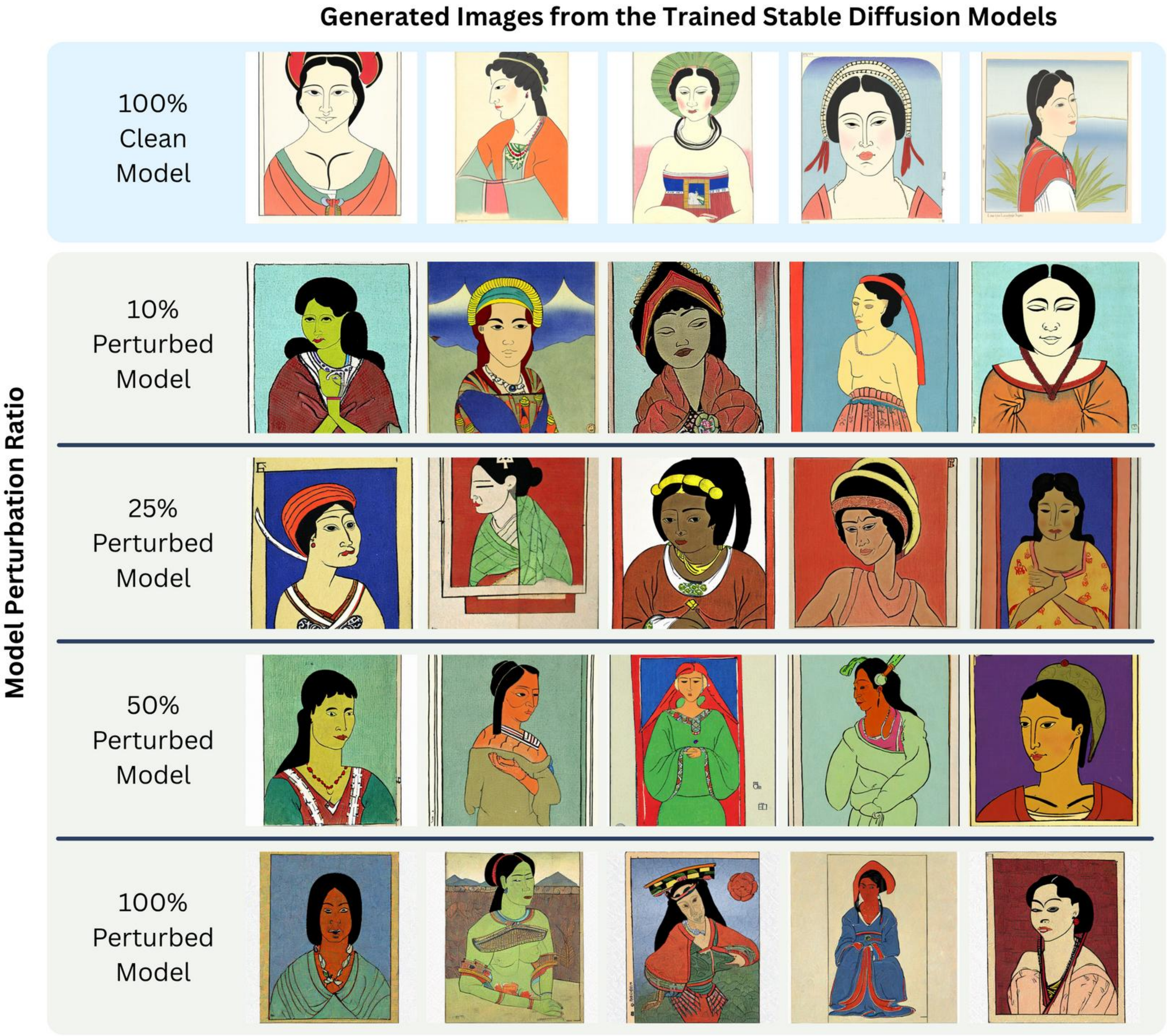
Output: **30 Perturbed/Protected Images**

Experiment Setup

Train a Stable Diffusion model using 30 perturbed images.

Experiment Setup

Batch	Composition	Description
1	100% Clean Dataset	Fully clean dataset
2	100% Perturbed Dataset	Fully perturbed dataset
3	90% Clean, 10% Perturbed Dataset	Mixed Dataset 1 – Low perturbation ratio
4	75% Clean, 25% Perturbed Dataset	Mixed Dataset 2 – Moderate perturbation ratio
5	50% Clean, 50% Perturbed Dataset	Mixed Dataset 3 – High perturbation ratio

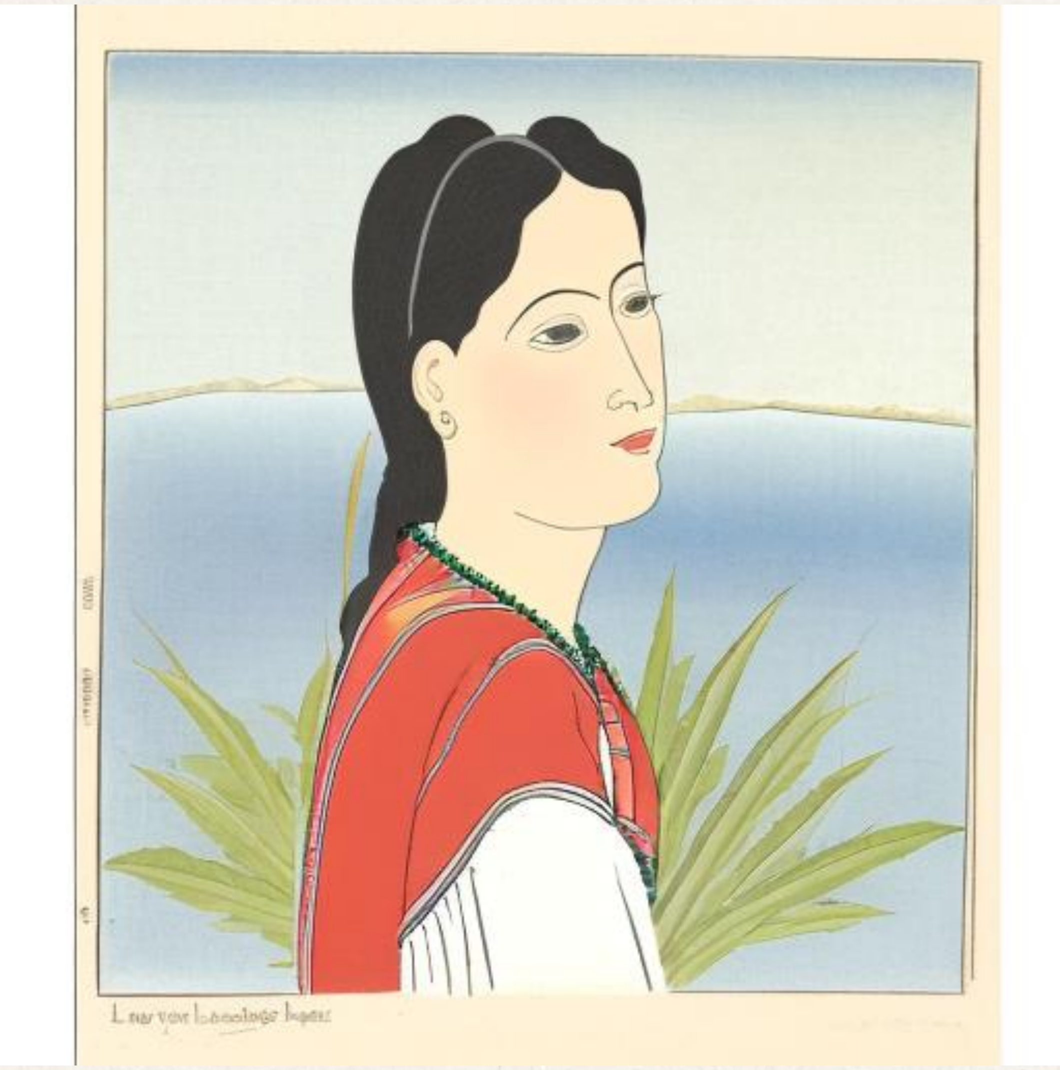


Model Outputs

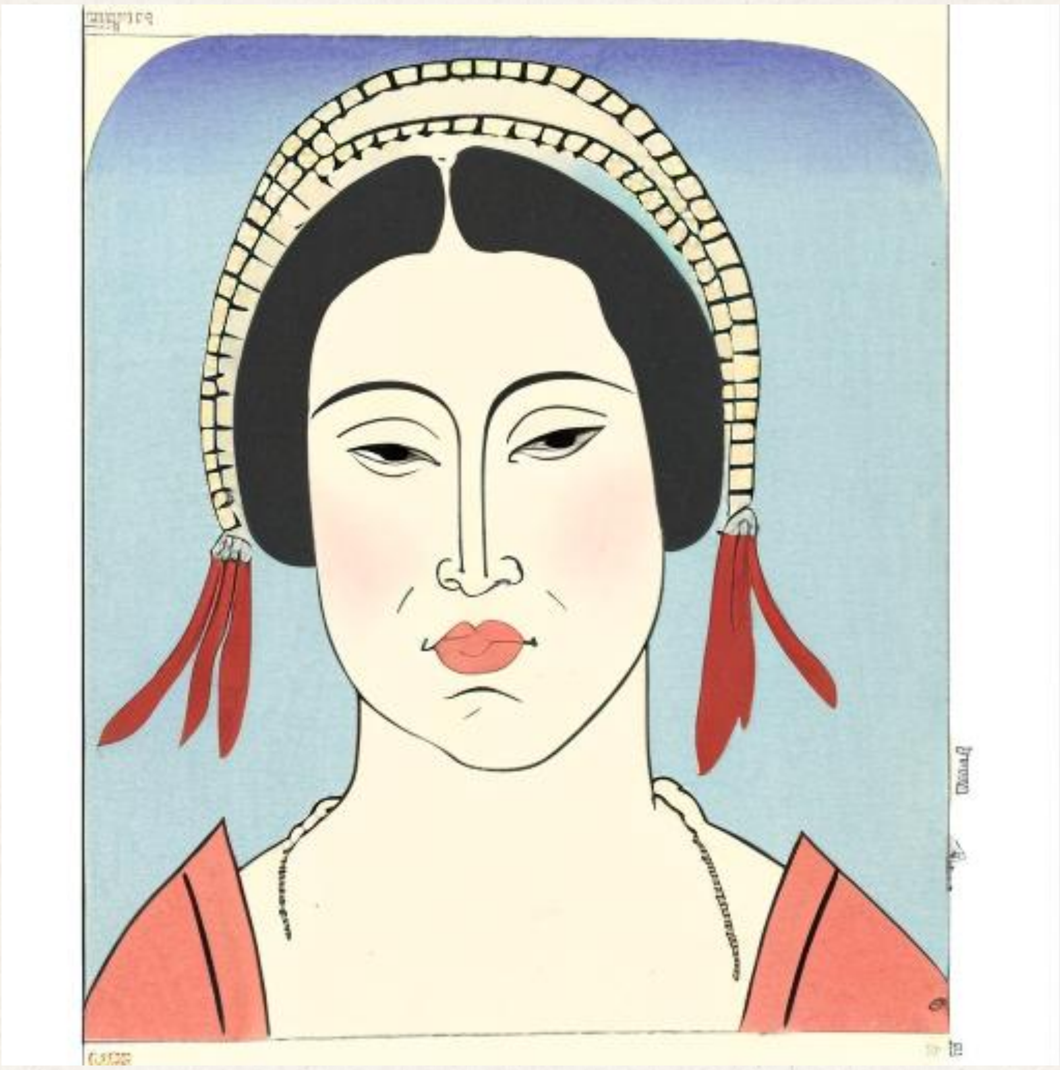
"an illustration of a woman in lauQui style"

100% Clean Model

ORIGINAL PAUL JACOULET PAINTING



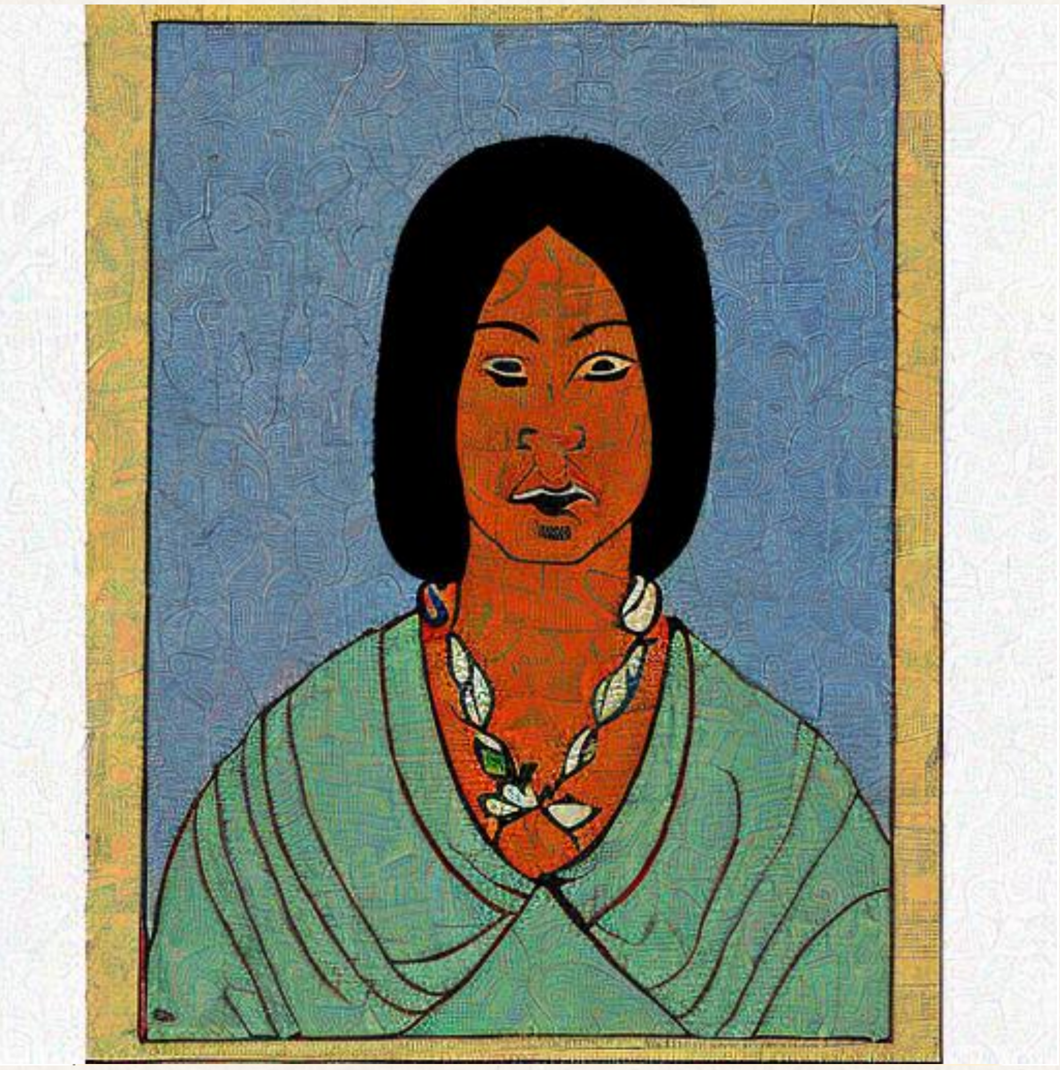
GENERATED IMAGE 1



GENERATED IMAGE 2

100% Perturbed Model

ORIGINAL PAUL JACOULET PAINTING



GENERATED IMAGE 1

GENERATED IMAGE 2

50% Perturbed Model

ORIGINAL PAUL JACOULET PAINTING

GENERATED IMAGE 1

GENERATED IMAGE 2

25% Perturbed Model

ORIGINAL PAUL JACOULET PAINTING

GENERATED IMAGE 1

GENERATED IMAGE 2

10%Perturbed Model

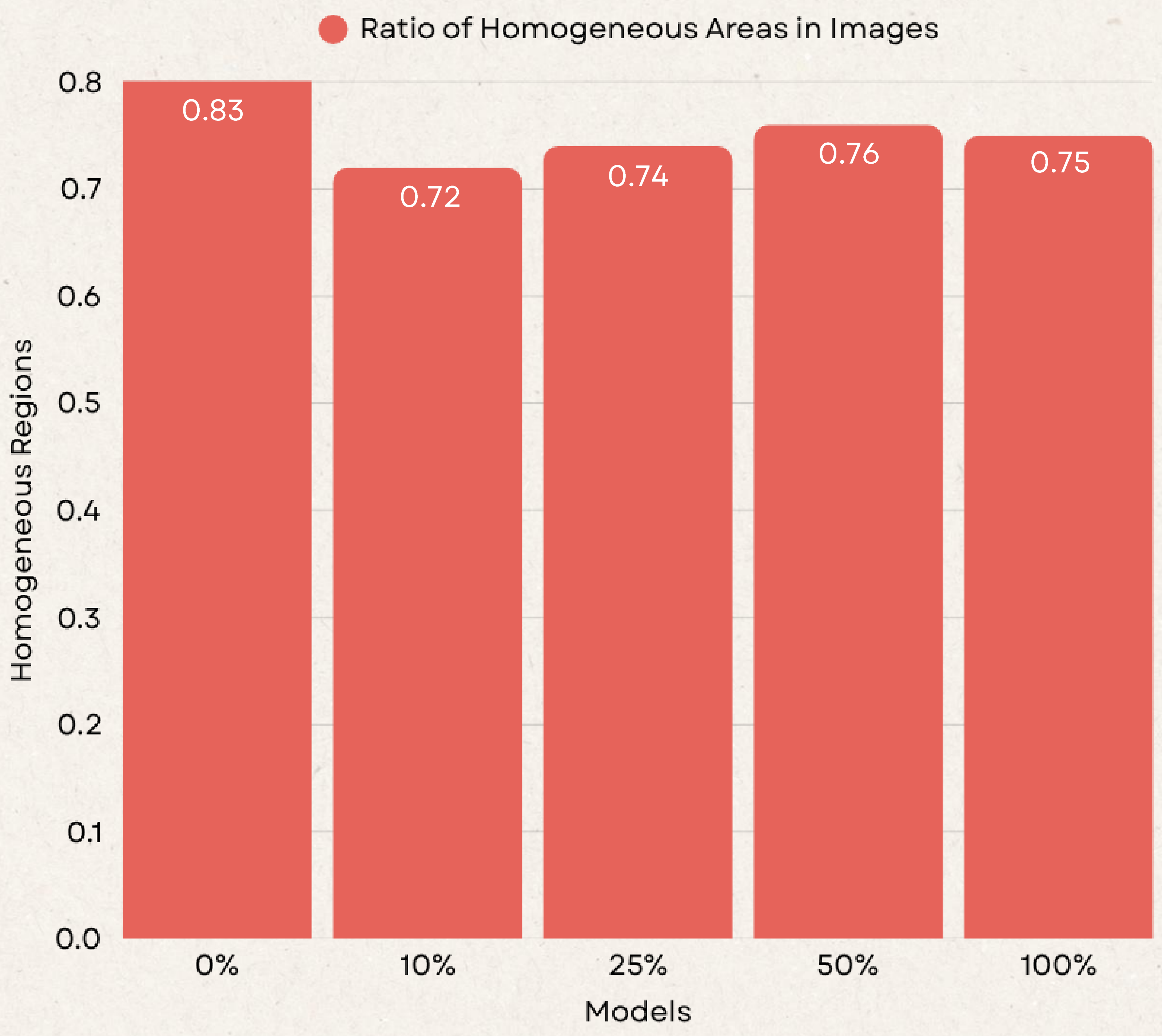
ORIGINAL PAUL JACOULET PAINTING

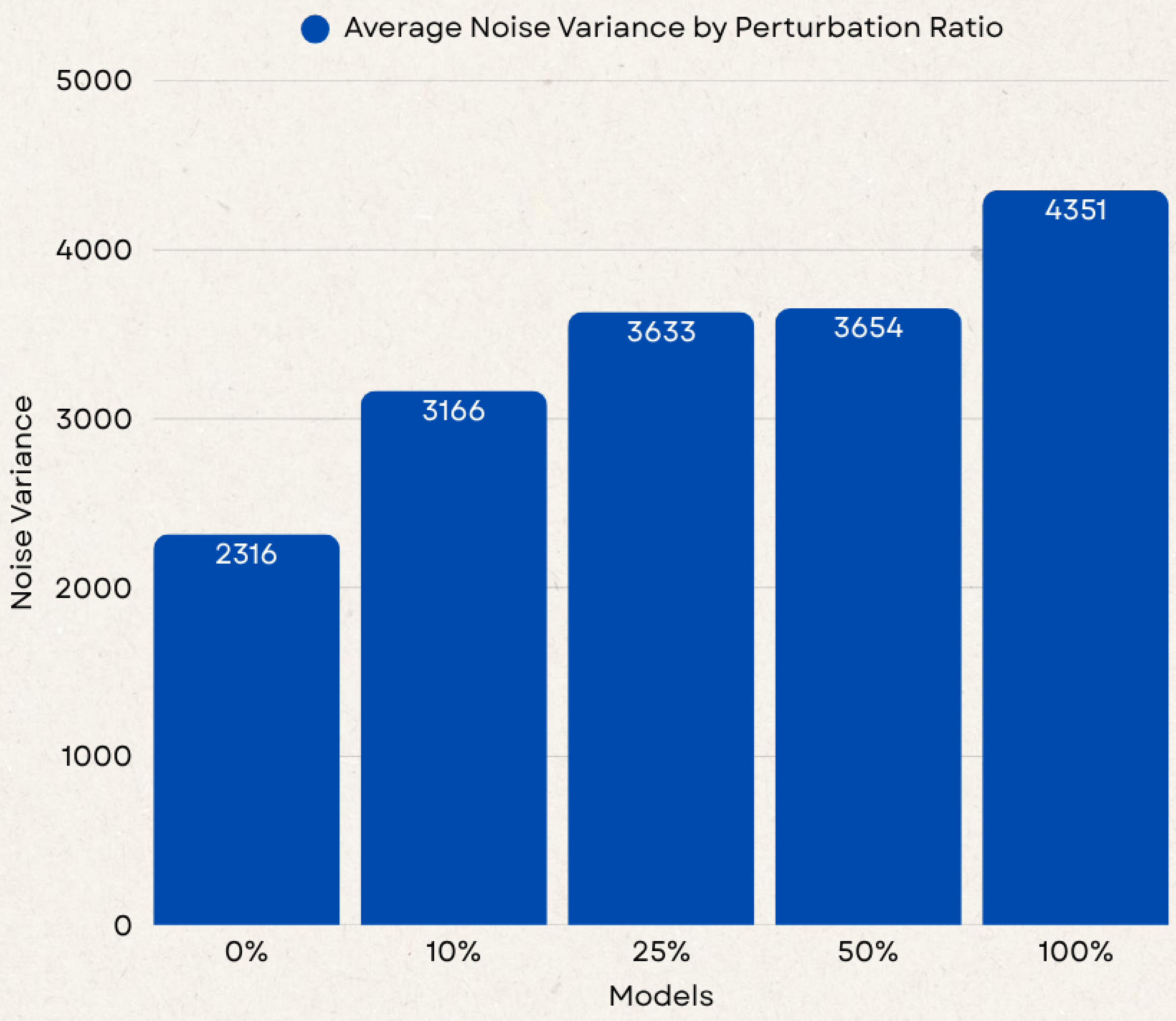
GENERATED IMAGE 1

GENERATED IMAGE 2

Evaluation

Noise Variance Distribution

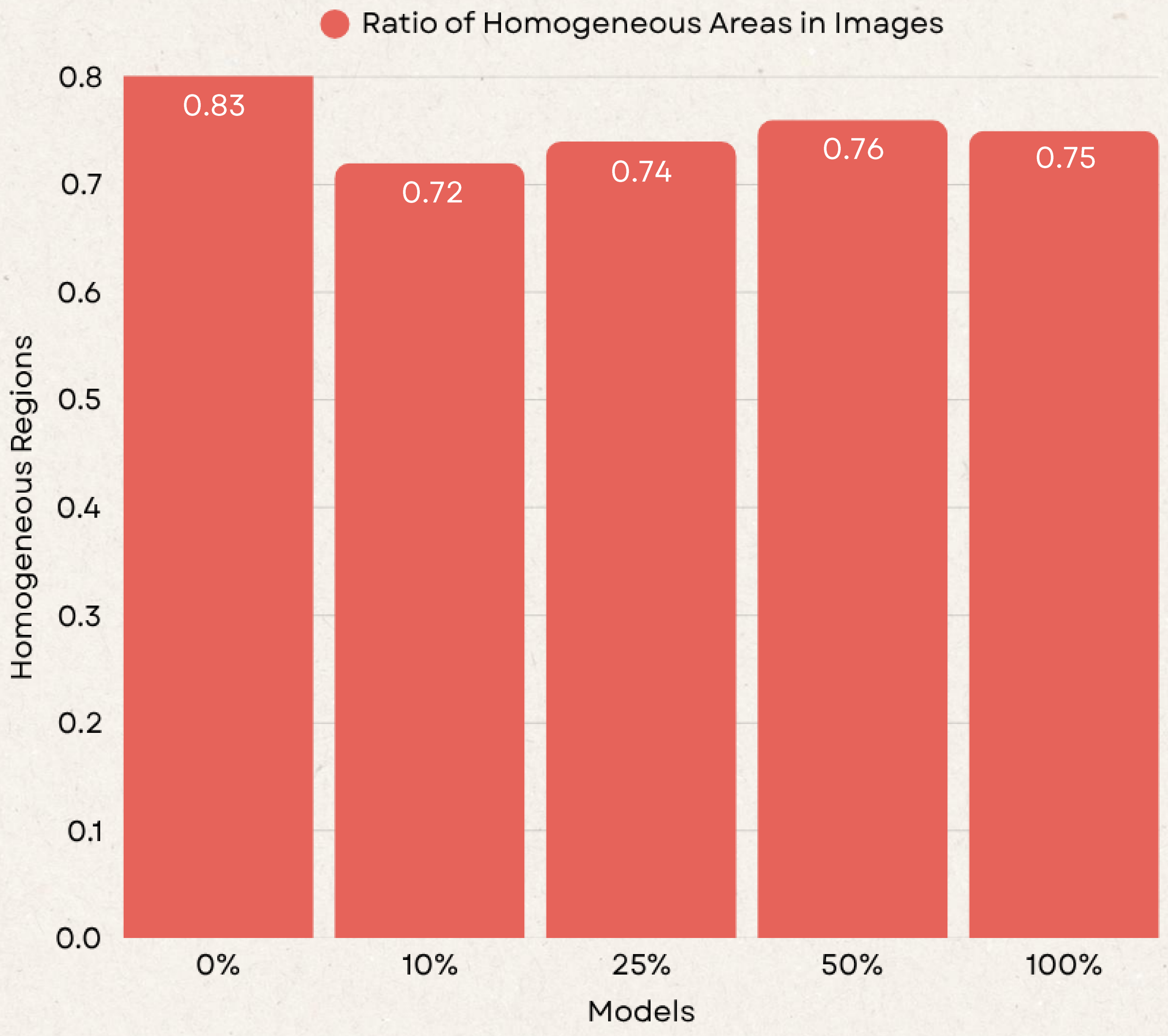


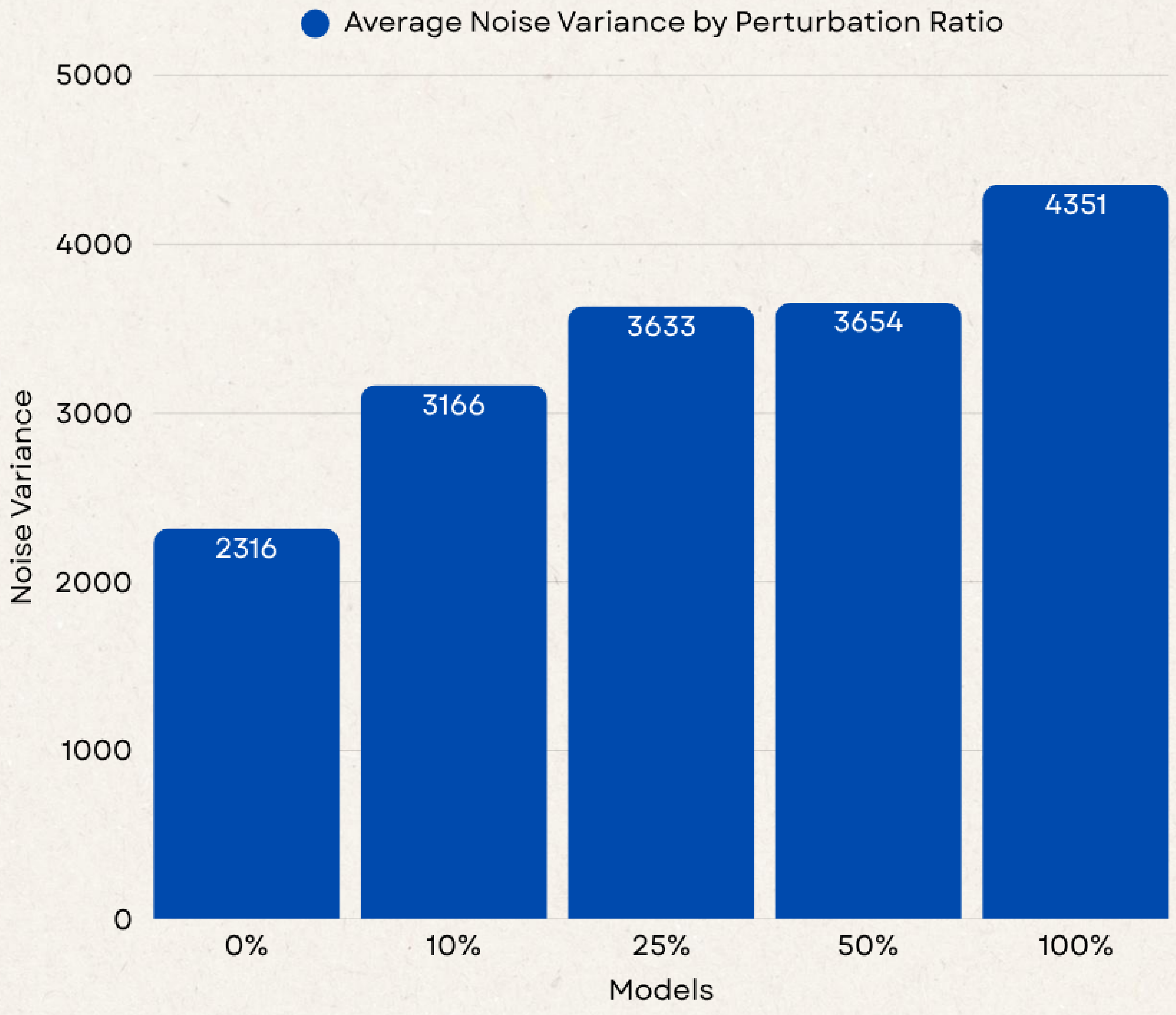


Evaluation

Noise Variance Distribution

100% perturbed = low homogeneous areas, highest noise.
100% clean = highest homogeneous areas, lowest noise.

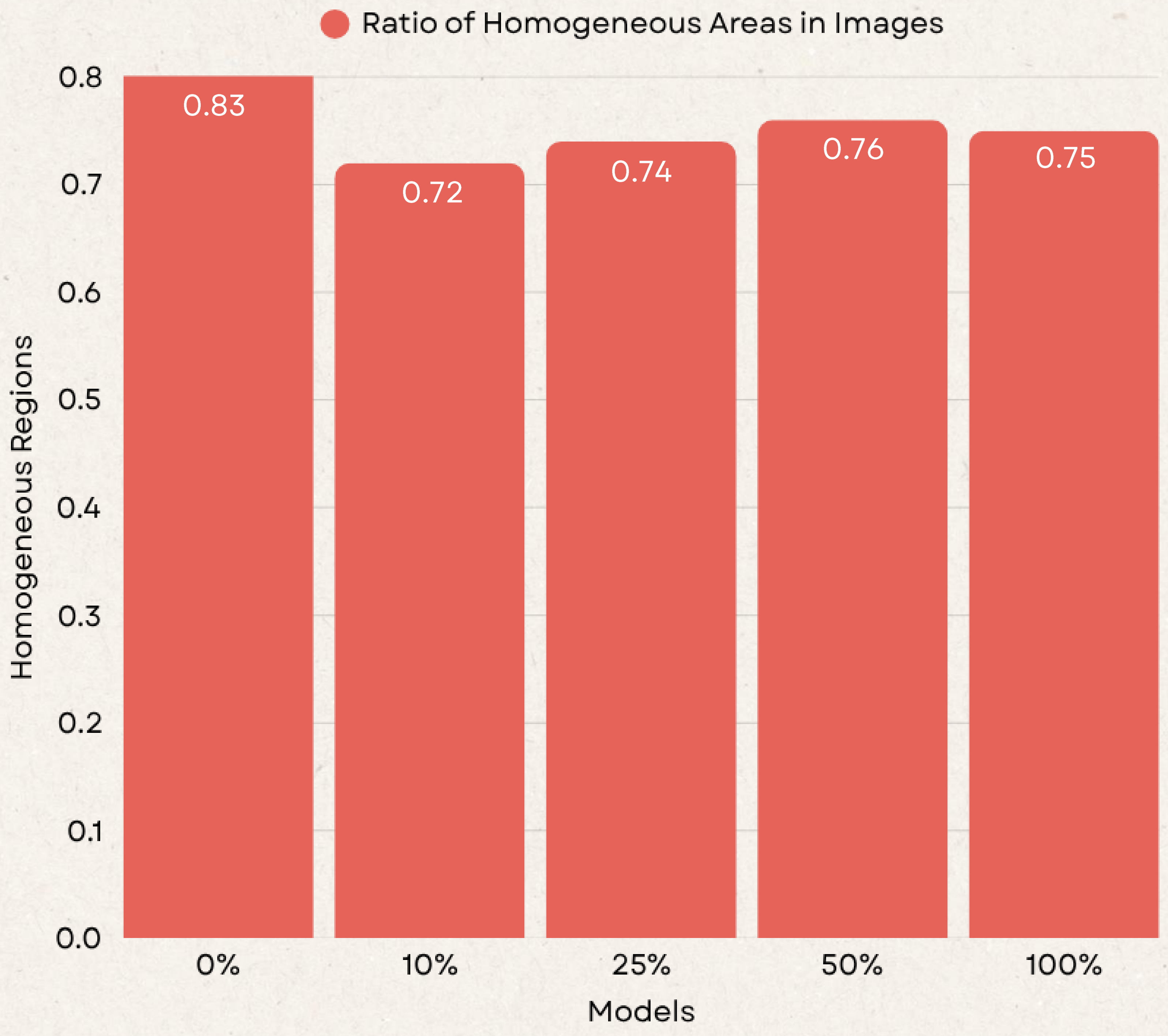


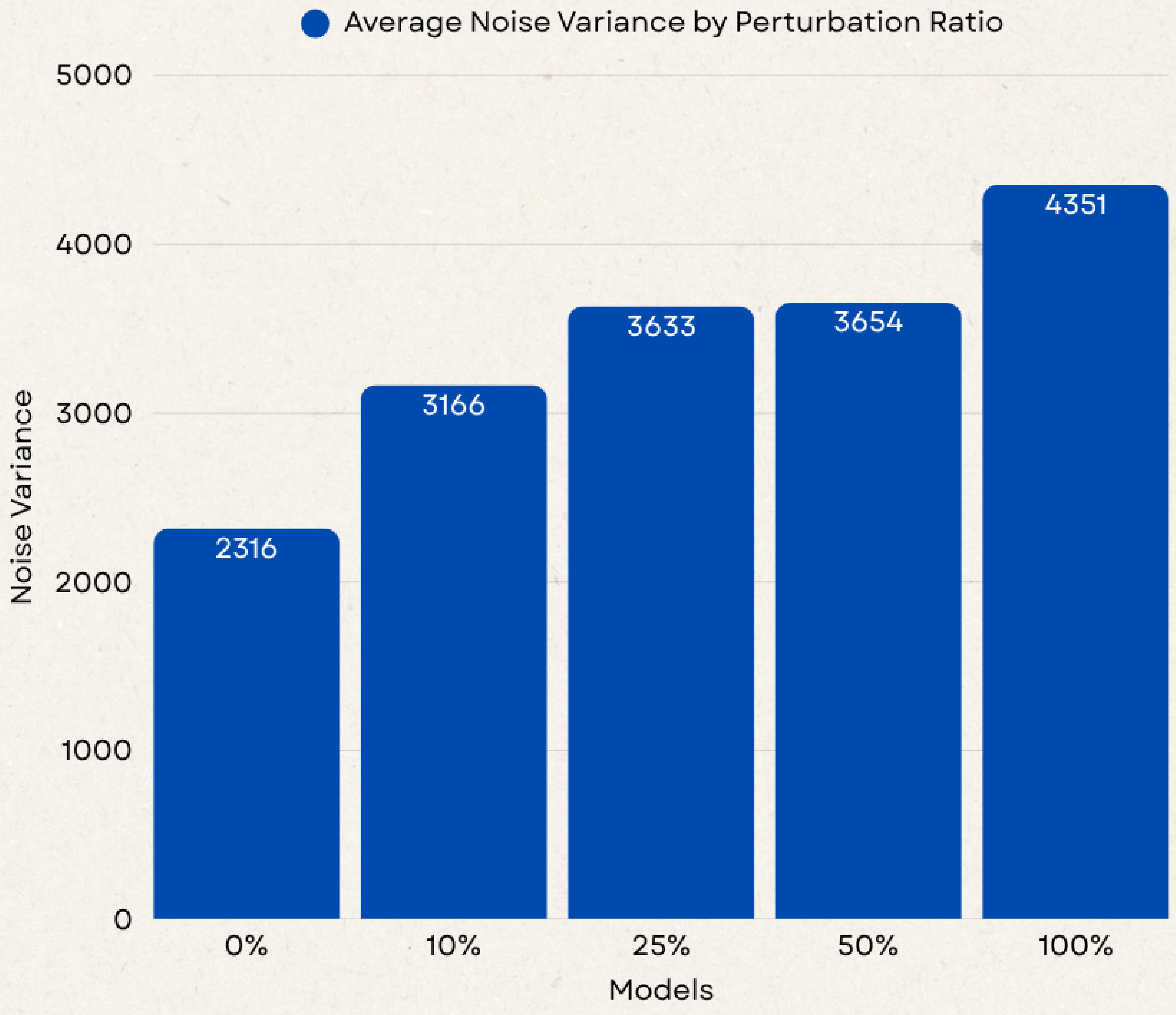


Evaluation

Noise Variance Distribution

Models trained on 10–50% perturbed datasets showed increasing noise variance with higher perturbation. All had significantly more noise than the clean dataset.





Therefore,

EVEN MINIMAL PERTURBATION—JUST 10% OF TRAINING IMAGES—SIGNIFICANTLY DISRUPTS DIFFUSION MODEL LEARNING.

Therefore,

THE MEMORY USAGE OF AINS DURING PERTURBATION IS WITHIN THE TARGET RANGE OF 4GIB OF VRAM, MAKING IT *MORE MEMORY EFFICIENT* THAN EXISTING PERTURBATION TOOLS.

Future Work

- EXTEND PROTECTION TO OTHER MODALITIES, SUCH AS IMAGE-TO-IMAGE MODELS.
- ADAPT TO NEWER MODEL VERSIONS (E.G., SD 3.5+).
- IMPROVE RESISTANCE TO IMAGE PURIFICATION TECHNIQUES (E.G., COMPRESSION, RESIZING).
- IMPROVE PERTURBATION IMPERCEPTIBILITY.

Thank you.