SOFTWARE ENGINEERING PROJECT

Vid-Scratch
(Project Proposal)

BY

Nantawan Paramapooti
Pichayoot Tanasinanan

DEPARTMENT OF COMPUTER ENGINEERING
FACULTY OF ENGINEERING
KASETSART UNIVERSITY

Academic Year 2026

Vid-Scratch
(Project Proposal)

BY

Nantawan Paramapooti
Pichayoot Tanasinanan

This Project Submitted in Partial Fulfillment of the Requirement for Bachelor Degree of Engineering (Software Engineering)
Department of Computer Engineering, Faculty of Engineering KASERTSART UNIVERSITY
Academic Year 2026

Approved By:

Advisor . . . . . **Date** . . . . . .
(Punpiti Piamsa-nga)

Co-Advisor . . . . . **Date** . . . . . .
(Prof. Chantana Chantrapornchai)

Head of Department . . . . . **Date** . . . . . .
(Punpiti Piamsa-nga)

Abstract

Currently, Generative AI has significantly advanced in producing high-quality video content, but it also introduces risks such as deepfake misuse, copyright infringement, and malicious manipulation. The Protractor project presents an AI-driven video poisoning processor designed to counteract the threats posed by Generative AI video models. By adding imperceptible perturbations to videos, Protractor ensures that while the video remains unchanged to the human eye, AI models misinterpret and degrade their outputs when trained on poisoned data.

The system leverages Breaking Temporal Consistency (BTC-UAP) and Spatially Transformed Adversarial Attacks (stAdv) to disrupt both frame-by-frame spatial details and motion-based temporal consistency, preventing AI from accurately learning patterns from poisoned videos. Additionally, the project implements adversarial noise embedding, perceptual similarity loss, and automated AI pipeline optimizations to maintain high fidelity for human viewers while corrupting AI training datasets.

The Protractor system is built for content creators, artists, and copyright holders who wish to protect their work from unauthorized AI training. Experimental results show that video poisoning significantly disrupts AI-generated outputs, making it a practical defense against AI exploitation and data misuse.

Keywords: Video Poisoning, Generative AI, Adversarial Attack, Deepfake Protection, AI Security

Acknowledgement

We would like to express our sincere gratitude to the Department of Computer Engineering, Kasetsart University for providing us with invaluable resources, technical knowledge, and support throughout the development of this project. Their guidance has been instrumental in shaping the Protractor system.

We are deeply thankful to our advisor, Assoc. Prof. Dr. Punpiti Piamsa-nga, for his expertise and mentorship in the field of image processing and adversarial attacks, which have been critical to the project's success. We also extend our heartfelt appreciation to Prof. Chantana Chantrapornchai, Ph.D., for her insightful guidance on parallel computing and optimization, enabling us to refine the efficiency of our AI pipeline.

Additionally, we would like to acknowledge the research communities and open-source contributors whose work on adversarial robustness, video processing, and AI security provided essential knowledge and tools that contributed to this project's implementation.

Finally, we would like to thank our faculty members, colleagues, and everyone who has supported and encouraged us throughout this journey. Their insights, discussions, and feedback have been invaluable in developing a meaningful and impactful solution to counter AI exploitation in video generation.

Nantawan Paramapooti
Pichayoot Tanasinanan

Table of Contents

List of Tables

Page

List of Figures

Chapter 1

Introduction

1.1 Background

Video Generative AI[0] is artificial intelligence models that create videos based on textual descriptions, images, or existing video inputs.

This AI has been utilized in many ways, such as to use AI character models to use the product like fashion and makeup to automate advertisement, aid in generating VFX[8], Enhance render quality with lower-end PC[12], Quickly upscale videos resolution, etc.

But the conception of Generative AI has been controversial in itself

It was trained with data scraped from the internet[4], with no regards to copyright permission or robot.txt[29] that ask them to not use their data on their website for AI training. Of course, this sparked an outrage amongst the owner of the stolen works, who rely on commission or unique products as their main income source. Video AI is no different. Some AI engineers have been transparent on sourcing their training data only from copyright-free and ethical datasets[3] (AnimateDiff, Adobe Firefly, Text2Video) , but the majority of the existing trained AI models[2] does not make the separation at all (MidJourney, Sora, CogVideoX, Runway, Kling, etc.) , to the point that they openly advertised fine-tuned models that generate artwork in the style of specific artists, using their names.

The widespread use of this AI has led to concerns about copyright infringement[1], Many artists and content creators advocate for stricter regulations to protect their works from being used without consent, to

protect the future of their career.

The controversy surrounding Generative AI isn't just about copyright, it's about how it devalues artistic labor. The issue isn't that artists can't produce higher-quality work, but that commissioners are willing to accept cheaper, lower-quality AI imitations that are 'good enough' for their purposes.

However, this lack of detail and cohesion will accumulate over time, as AI operates on token-based prompts rather than true creative intent. This shift doesn't just harm individual artists, it weakens the entire support system and foundation of the creative industry. As demand for human-made art declines, fewer aspiring artists will have opportunities to develop their skills, leading to a dwindling talent pool. Without skilled artists, industries that rely on craftsmanship such as animation, illustration, and game design will suffer, ultimately lowering the overall quality of creative work in the long run.

Once that foundation is lost, rebuilding it becomes nearly impossible. Just as Disney can no longer return to the same level of hand-drawn 2D animation it once mastered after it shifted entirely to 3D animation, an overreliance on AI risks permanently degrading the artistic standards and craftsmanship that define creative industries. Without a strong base of skilled artists, industries like animation, illustration, and game design will struggle to maintain the depth and quality that set them apart in the first place.

Consumers have already expressed their disappointment with the rise of AI-generated slop content, which has flooded media platforms at an overwhelming speed, pushing high-quality, human-made work out of recommendation algorithms. However, beyond AI itself, grifters[10] have exploited this shift by infringing on copyrighted works, fine-tuning[39] models on stolen art, and mass-producing low-effort imitations for profit.

These bad actors don't just flood platforms with AI-generated content, they actively leech off the success of original creators, diverting attention, funds, and opportunities that should have gone to real artists. As their stolen content dominates search results and recommendation feeds, genuine creators struggle to gain visibility, leading many to abandon the platform entirely.

As AI-generated content continues to dominate social networks, we move closer to the so-called 'Dead Internet Theory'[11] where authentic human interaction is drowned out by automated, mass-produced content. The internet, once a space for organic connection and creativity, is becoming increasingly hostile due to rampant exploitation, with grifters[10] and engagement farmers prioritizing profit over meaningful discourse. If this trend continues, the internet risks becoming a hollow, artificially-generated echo chamber, devoid of genuine human expression.

Lastly, One of the most alarming consequences of AI-generated video technology is impersonation[5], often referred to as "deepfakes[6]." AI can create realistic videos of individuals, making them appear to say or do things they never did. This poses risks in identity theft[7], misinformation, fraud, and political manipulation[9]. The ability to create hyper-realistic fake videos raises concerns about trust in digital content and calls for advanced detection methods to counteract malicious use.

Data poisoning is a method of corrupting AI models by injecting misleading or harmful data into their training sets[13], ensuring that generative models[14] cannot easily exploit original artistic content. It is an aspect of data security[15], and restrict malicious actors from exploiting your data against your interests.

1.2 Problem Statement

The problem this project aims to solve has been well defined in the background section.

As of currently, there is a solid foundation of research on the adversarial poisoning tactics[40] against Video Generative AI, but there is currently no software that simplifies the process for common use. People may have the idea to extract the video's graphics frame by frame, then use the existing static Image Poisoning Processor to Poison frame by frame, then reassemble them back into their video form, but there's still technical problems, such as:

1. Manually poisoning frame by frame is inconvenient for production use.
2. Existing solution's processing time scales horribly with video duration and fps. A 10 second video with 30 fps could take as long as 8 hours with default parameters on Glaze, and that is on the premise that the app doesn't crash as it can only handle up to 70 images queued to poison maximum.
3. Static Image poisoning tactics are less effective against Video generative AI.

1.3 Solution Overview

This software seeks to simplify the process of video poisoning to be easy to use. The software will only need the user to input their video, set some preferences, start the process and wait for the poisoned video output in their designated folder. While being effective against generative AI and efficiently optimizing hardware resources to process larger video; ranging from 5 minutes to 2 hours, to be processed fast and reliable enough for target users such as filmmakers, content creators, and studios to incorporate this in their workflow.

1.3.1 Features

1. Video Poisoning: Poison the video by injecting perturbations[41] that tricks the AI into learning false patterns and ruins its output consistency and quality. Before

you may start, you'll have to input your video, select the output folder, then start the poisoning process[36].

2. Adjust Poisoning Parameters[15] Settings: Set predefined Parameters such as perturbation weights[16] or output quality[17] to set the perturbation strength[18] and output quality[17]. More parameters may be added depending on the available parameters of the system's poisoning methods[19].

3. Video format support: Input and Output only supports .mp4 for this project, but may add more file format supports in the future.

4. Hardware optimization: Optimize the available hardware to minimize processing time duration. This would be done automatically but may allow users to set hardware themselves if deemed appropriate.

1.4 Target User

- **Digital Content Creators & Video Artists[20]:** They have had their creations[36] used as training data[21] without their permission to replicate their work, making their creative, unique, curated work being buried amongst their AI copies that hurt their profits[22] and fame.

- **Industry Professionals in Media & Entertainment[23]:** Animation studios are at risk of having their creative works being exploited to create lower quality but faster animations. This could result in the death of the Animation industry[24] altogether as Animator and other creatives being laid off after their works had been trained on AI and the audience ends up with an incoherent meaningless repetitive mediocre slop[25] because the company thought that was good enough for the audience and artist become more distrustful of sharing their works online.

- **Anti-AI social media platforms:** Cara, BlueSky, Teezr, VGen are against any AI-generated content[26] on their platforms. This could be part of their feature to protect their userbase's video against being used to train on AI.
- **Individuals who do not want their videos to be used to train generative AI:** From the dangers of deepfakes[6], regular people do not want their face to be used to train generative AI in General, but data scraping was done without considering their consent. This will force data scrapers[27] to exclude poisoned data[28] from their training dataset[3].

1.5 Benefit

The Protractor system protects video content from non-consensual AI training[30] by applying adversarial techniques[31] that disrupt AI perception while remaining imperceptible[32] to humans.

- It breaks AI generated video quality and frame consistency[33], stopping deepfake from producing similar creations[34].
- It enhances intellectual property[35] protection for creators and safeguards the creative industry from AI-driven content theft[36].
- Its easy-to-use implementation allows everyone to apply AI poisoning[37] without requiring advanced technical expertise.

1.6 Timeline

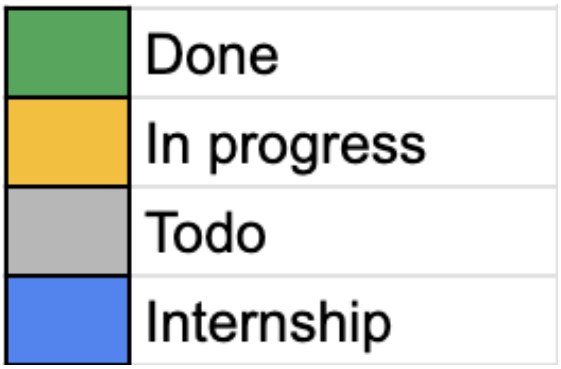

Figure 1.1: Color Meaning chart

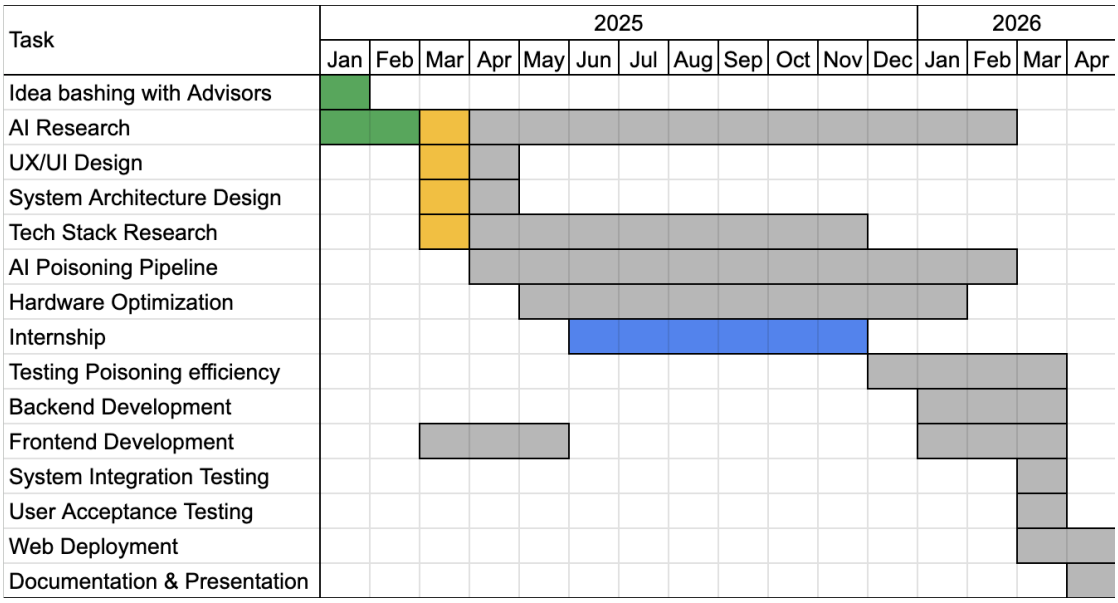

Figure 1.2: Timeline of Project development

In Figure 1.2. is the current timeline of our project plan. You may look up to Figure 1.1. for color code meanings.

1.7 Terminology

[0]AI : artificial intelligence
[1]copyright infringement : violating copyright law over a content
[2]AI models : AI programs consisting of complex mathematical and computational techniques to process vast amounts of data and extract meaningful insights.
[3]dataset : collections of data used to train AI models.
[4]scraped from the internet : automatically collecting data from online sources, often using web crawlers or scrapers.
[5]Impersonation : The act of fraudulently imitating a person to deceive others.
[6]deepfakes : AI-generated videos that convincingly replace a person's likeness or voice with another, often for deceptive purposes.
[7]identity theft : The unauthorized use of someone's personal information to commit fraud or other crimes.
[8]VFX : Computer-generated effects used in media, often automated by AI. [9]political manipulation : The use of deceptive tactics, such as deepfakes or AI-generated propaganda, to influence public opinion or elections.
[10]Grifters : People who try to get you in get-rich-quick schemes that turned out to be a total waste of time.
[11]Dead internet theory : A conspiracy introduced by IlluminatiPirate on the forum Agora Road's Macintosh Cafe esoteric board. Referring to the future where genuine human interaction is overtaken by bots and AI generated content due to the sheer amount and available.
[12]lower-end PC : A computer with limited computational power, struggling with high-end tasks. [13]poisoning tactics : The tactics of poisoning a graphics content that break AI when it trained on the poisoned piece of media content
[14]static Image poisoning processor : Refer to a program that adds "AI poison" to the input non-moving image.
[15]Poisoning : The process of 'poisoning' the input to make it break AI models when trained on, which will increase with the percentage of poisoned works in the dataset.

[16]perturbation weights : perturbation is added via a formula (x + x' = p; x is the original input, x' is the perturbation and p is the poisoned output, x' could be w*noise where w is weight and noise is the graphic of a randomized RGB image designed to make AI perform worse through computer vision) and added to the original image through the RGB channel of the original image. [17]output quality : The quality of the output after the input had been poisoned
[18]perturbation strength : How obvious the perturbation is in the poisoned output
[19]poisoning methods : methods to 'poison' an image
[20]Video Artists : Any artist that create video content, like animators or illustrator art timelapse where they post the process of creating their art
[21]training data : data that AI trains on
[22]profits : For the owner of the video, they may get their profits through commissions, platform revenue, merchandise, etc. Their profits are hurt because an AI copy could steal their originality, hard work or recommendation spots that would pay them.
[23]Industry Professionals in Media and Entertainment : Refer to any creatives who work in the Media and Entertainment industry.
[24]the death of the Animation industry : As the animation industry's jobs become unstable and at risk of being replaced by AI, either the next generation of workers have to sacrifice their limited resources to compete with the availability and speed (but lack of quality) of AI, or perish. As their investors and customers use AI instead for cheaper, faster work. Jobs that could be the transition role for newbies to developing the skills of a professional are being replaced by AI, which means that there's going to be less to no senior professional to pass the job on.
[25]slop : low quality content that's mediocre at best, but usually not good enough to provide any meaningful value to the consumer.
[26]AI-generated content : Content created from AI generation via a prompt or an input image
[27]data scrapers : Refer to entity that perform data scraping to collect data for any use
[28]poisoned data : data that has been 'poisoned' that will break the AI when it was trained on.

[29]robot.txt : A file that restricts web crawlers from accessing certain parts of a site. [30]non-consensual AI training : Refer to how AI trains on data without the data's owner consent.
[31]adversarial techniques : A data poisoning tactic where they change the data material to encourage AI to learn false patterns during backpropagation, while maintaining the perceptual similarity to the original work.
[32]imperceptible : Undetectable with the human eye
[33]frame consistency : How video graphics make sense between the previous, current and next frame. Low frame consistency means the video is flick-ery and objects and details appear and disappear more unpredictably.
[34]similar creations : Creative products that looks similar in style or appearance
[35]intellectual property : Legal rights that protect creations of the mind, such as art, music, inventions, patents, trademarks, and copyrights.
[36]AI-driven content theft : Unauthorized use or replication of copyrighted materials by AI systems. It is a copyright infringement
[37]AI poisoning : AI poisoning (also known as data poisoning) is a method used to corrupt or manipulate machine learning models by introducing misleading or harmful data.
[38]poisoning process : The process of 'poisoning' the input against AI
[39]fine-tuning : Adjusting a pre-trained model with specific data to specialize it. [40]adversarial poisoning tactics : Manipulating training data to disrupt AI performance. [41]perturbations : Small changes or modifications made to data to affect the behavior of a model, often used in adversarial attacks to mislead or deceive AI systems.

Chapter 2

Literature Review and Related Work

Generative AI has seen rapid advancements, particularly in video synthesis and manipulation. However, as AI-generated content becomes more sophisticated, concerns over deepfake misuse, copyright violations, and data exploitation has risen. Several existing tools attempt to address these challenges, but they focus primarily on static image protection rather than video poisoning.

2.1 Competitor Analysis

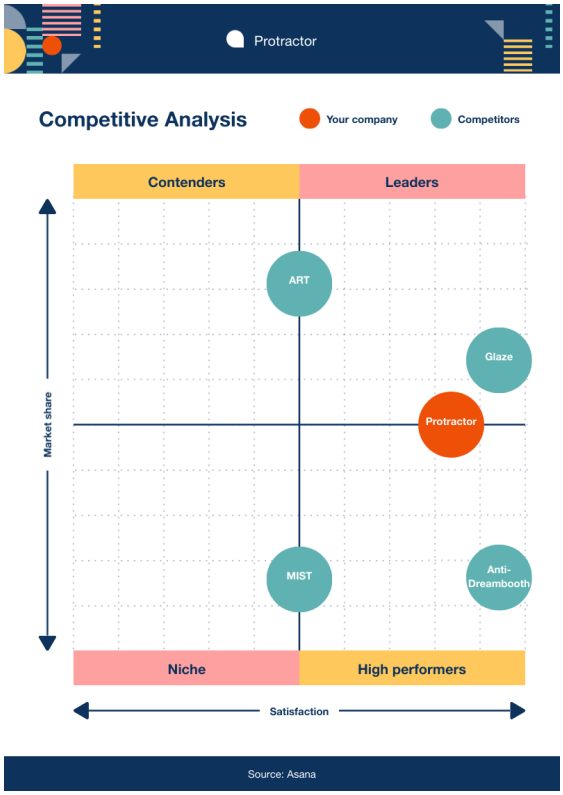

Figure 2.1: Competitive Landscape

Because there are currently no poisoning processors for videos, our current competitor would be static image poisoning processors, in a scenario where they extract the video's graphics frame by frame, and poison each of them as a static image, and then reassemble them back as a video.

The goal of the software is to protect the video from AI training by data poisoning method Adversarial attack; to break AI's accuracy by adding information to the data that is imperceptible to the human eye.

1. Glaze

   Glaze is a tool that applies adversarial perturbations to digital artwork to prevent AI from learning its style. It modifies the image in a way that disrupts AI training while keeping the visual impact minimal to the human eye. Like other competitors, it only targets static images. Their poisoning method is less effective against AI due to video denoiser Autoencoder using more techniques such as temporal Consistency or attention seeking transformer methods which makes perturbations for static images less effective, and takes too long to be implemented Frame by frame, as a 10 minutes long video with 30 fps could take a minimum of 8 hours.

2. MIST

   MIST is another adversarial tool designed to mislead AI models into misinterpreting images, effectively reducing the accuracy of AI-based recognition and training systems. It faces the same challenges as Glaze for its implementation in video poisoning.

3. Anti-Dreambooth

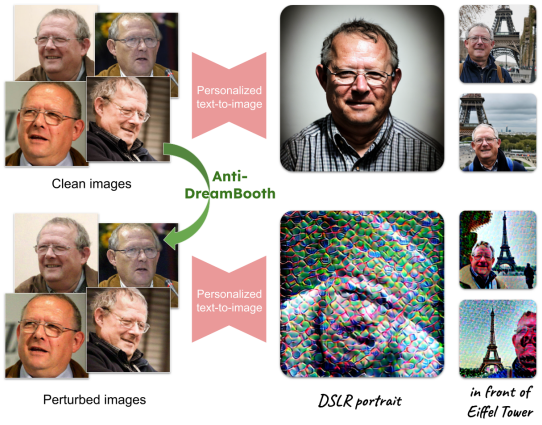

Figure 2.2: Anti-DreamBooth sample demonstration

Anti-Dreambooth targets models that fine-tune on small datasets, introducing noise patterns that degrade the ability of AI to generate imitative outputs. It faces the same challenges as Glaze for its implementation in video poisoning.

4. ART : Adversarial Robustness toolbox

   The Adversarial Robustness Toolbox (ART) is a broader security-focused framework that provides methods for generating adversarial attacks and defenses against AI-based recognition. There are currently no video perturbation methods implemented yet.

2.2 Literature Review

1. stAdv

   stAdv : Spatially Transformed Adversarial Attack

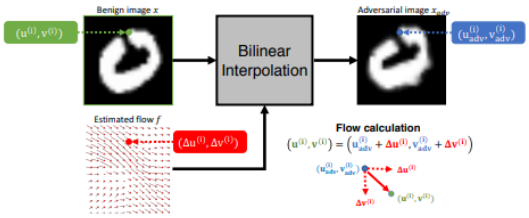

Figure 2.3: Explanation of how perturbation is added

   stAdv is a type of adversarial attack based on local geometric transformations.

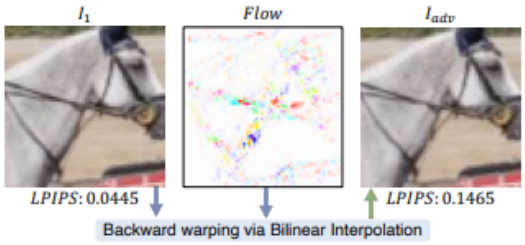

Figure 2.4: Explanation of how perturbation is added with LPIPS to measure perceptual similarity between the original and the poisoned output

   It has been shown by CAM attention visualization to shift the focal point of attention away from the original's, and proven to be stronger than FGSM and C&W method

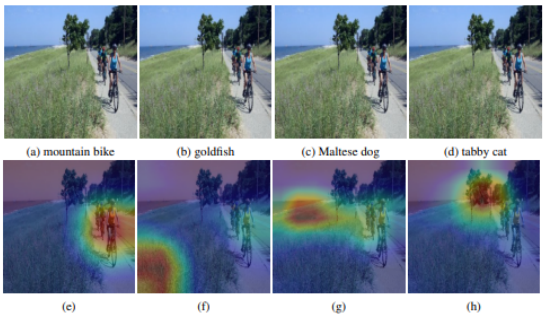

Figure 8: CAM attention visualization for ImageNet inception_v3 model. (a) the original image and (b)-(d) are stAdv adversarial examples targeting different classes. Row 2 shows the attention visualization for the corresponding images above.

Figure 2.5: CAM attention results after image had been poisoned

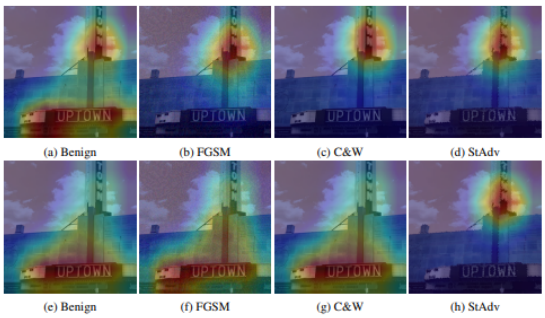

Figure 9: CAM attention visualization for ImageNet inception_v3 model. Column 1 shows the CAM map corresponding to the original image. Column 2-4 show the adversarial examples generated by different methods. The visualizations are drawn for Row 1: inception_v3 model; Row 2: (robust) adversarial trained inception_v3 model. (a) and (e)-(g) are labeled as the ground truth "cinema", while (b)-(d) and (h) are labeled as the adversarial target "missile."

Figure 2.6: CAM attention results of stAdv compared to FGSM and C&W perturbation where Benign is the control sample

2. BTC-UAP

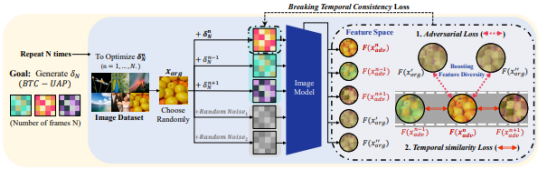

Figure 2.7: BTC-UAP poisoning logic

BTC-UAP : Breaking Temporal Consistency - Universal Adversarial Perturbation

BTC-UAP is a technique that changes videos in a way that confuses AI, making it harder to recognize patterns. This is crucial in staying subtle enough so the output image stays perceptually similar to human and not easily removable by Video Denoising Autoencoder, but strong enough to break AI's Temporal Consistency after they trained on data poisoned with BTC-UAP.

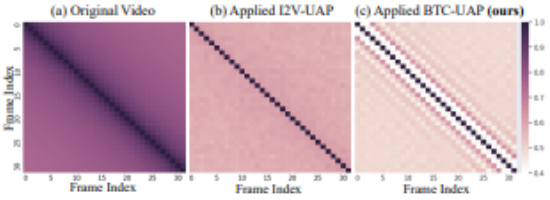

Figure 3: **The feature similarity of frames within videos.** This heatmap shows the average feature similarity between frames in the UCF-101 dataset, with brighter colors indicating lower levels of similarity.

Figure 2.8: Heatmap of Temporal Consistency; darker means better Consistency

Here it is demonstrated that the Temporal Consistency of BTC-UAP is the best perturbation that breaks AI's performance worst, while the output is still perceptually identical to the original video input.

To optimize our software and be a reliable data security tool, This project will be built upon various researches like these to ensure their effectiveness against bad actors from using the ever-evolving State-of-the-Art Generative AIs as a tool to exploit user's data.

Chapter 3

Requirement Analysis

3.1 Stakeholder Analysis

1. **Animators & Video Artists:** The primary stakeholders are content creators and video artists who want to protect their work from unauthorized AI training.
2. **Regulatory & Legal Bodies:** Government and industry groups ensuring compliance with copyright laws, AI ethics, and intellectual property protection.
3. **Cybersecurity & Digital Rights Organizations:** Groups that focus on copyright protection, AI misuse prevention, and ethical AI practices. They may promote or validate the effectiveness of Protractor.

3.2 User Stories

3.2.1 Upload & Protecting Video Content from AI Training

As a digital content creator or video artist,
I want to apply adversarial techniques to my videos to prevent AI models from copying my style,
so that I can protect my work from unauthorized AI training while keeping the video unchanged for humans.

3.2.2 Ensuring Video Quality for Human Viewers

As a digital content creator or video artist,
I want to apply AI poisoning techniques without affecting the

video's visual quality for humans,
so that my audience can still enjoy my videos without noticeable distortions.

3.2.3 Customizing Poisoning Parameters

As a digital content creator or video artist,
I want to adjust the poisoning settings such as intensity level and render quality before starting the process,
so that I can control how strongly the video is protected while balancing quality and performance.

3.2.4 Processing Videos Efficiently

As a content creator with large video files,
I want to have my videos processed quickly without long waiting times,
so that I can protect my videos without delays affecting my workflow.

3.2.5 Monitoring Poisoning Progress

As a user waiting for video processing to complete,
I want to see real-time updates on the status of my video poisoning,
so that I know how long the process will take and when my video will be ready.

3.3 Use Case Diagram

- ***User***—A default group of users which has all the basic functionalities
- ***User***—A default group of users which has all the basic functionalities

3.4 Use Case Model

3.4.1 Use Case 1: Uploading a Video

Actors: Pong (Digital Content Creator), Protractor (System)
Description: Pong wants to upload a video file to the system in preparation for poisoning.
Scenario:

1. Pong opens the Protractor web application.
2. System displays the upload interface.
3. Pong selects a supported video format file and clicks the upload button.
4. System validates the file and stores it for processing.

Alternative Flow: If the file type is unsupported, the system displays an error message.

3.4.2 Use Case 2: Setting Poisoning Parameters

Actors: Pong (Digital Content Creator), Protractor (System)
Description: Pong customizes the poisoning settings before starting the video poisoning process.
Scenario:

1. Pong navigates to the settings panel after uploading a video.
2. System displays available poisoning parameters.
3. Pong selects the settings.
4. System saves the selected configuration for use in the next step.

3.4.3 Use Case 3: Starting the Poisoning Process

Actors: Pong (Digital Content Creator), Protractor (System)

Description: After configuring the parameters, Pong starts the video poisoning process.

Scenario:

1. Pong clicks the "Start Poisoning" button.
2. System begins poisoning the video.
3. System displays a progress indicator or loading bar.

3.4.4 Use Case 4: Viewing the Poisoning Progress

Actors: Pong (Digital Content Creator), Protractor (System)

Description: Pong wants to track the progress of the poisoning process.

Scenario:

1. System updates a progress are processed.
2. Pong monitors the process status.
3. Once complete, the system notifies Pong that the video is ready.

3.4.5 Use Case 5: Downloading the Poisoned Video

Actors: Pong (Digital Content Creator), Protractor (System)

Description: Pong downloads the final poisoned video file to their device.

Scenario:

1. The system displays a "Download" button once poisoning is complete.
2. Pong clicks the button to download the poisoned video.
3. System sends the processed file to Pong's device.

3.5 User Interface Design

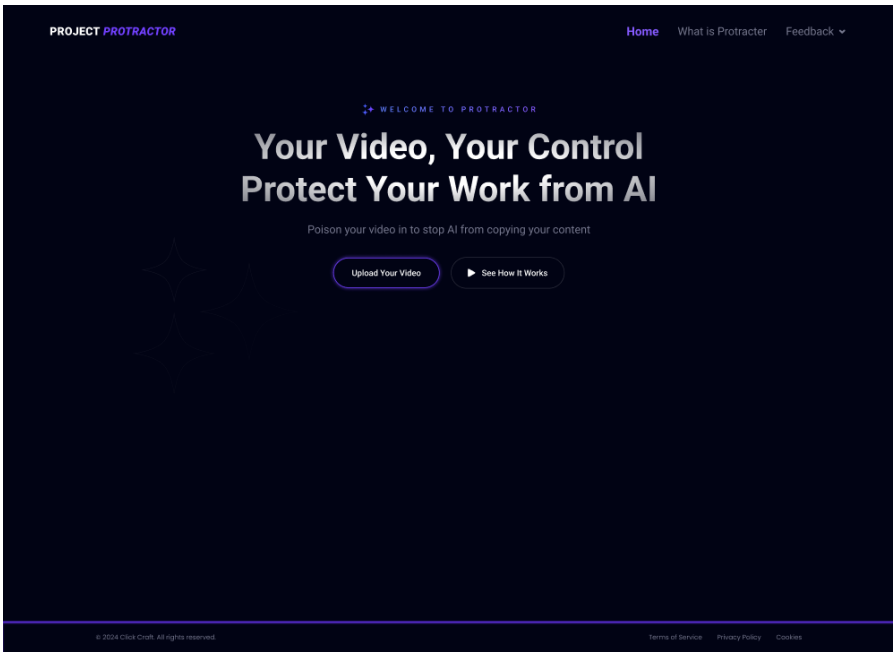

Figure 3.1: Mockup of Home Design

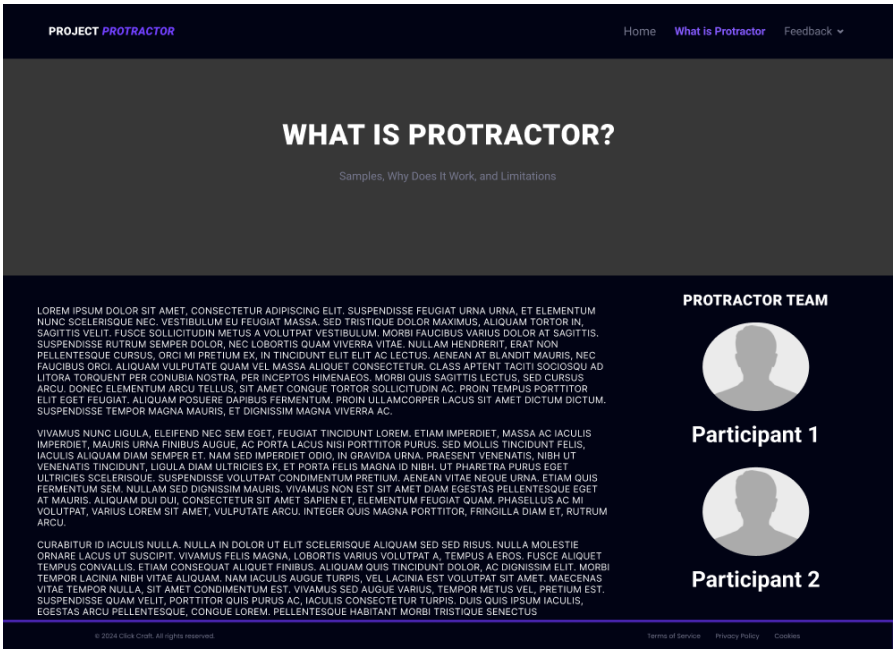

Figure 3.2: Mockup of What is Design

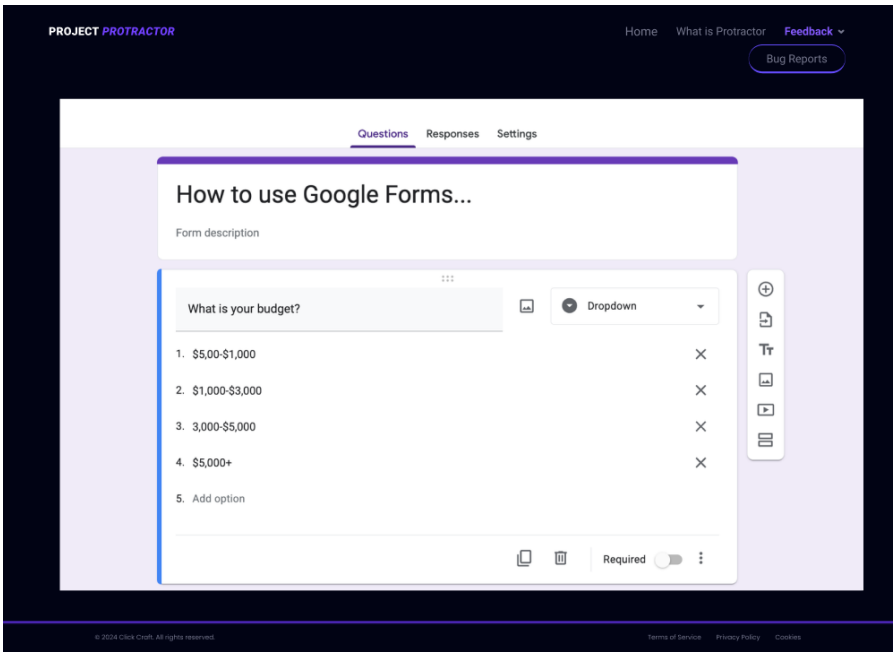

Figure 3.3: Mockup of Feedback Design

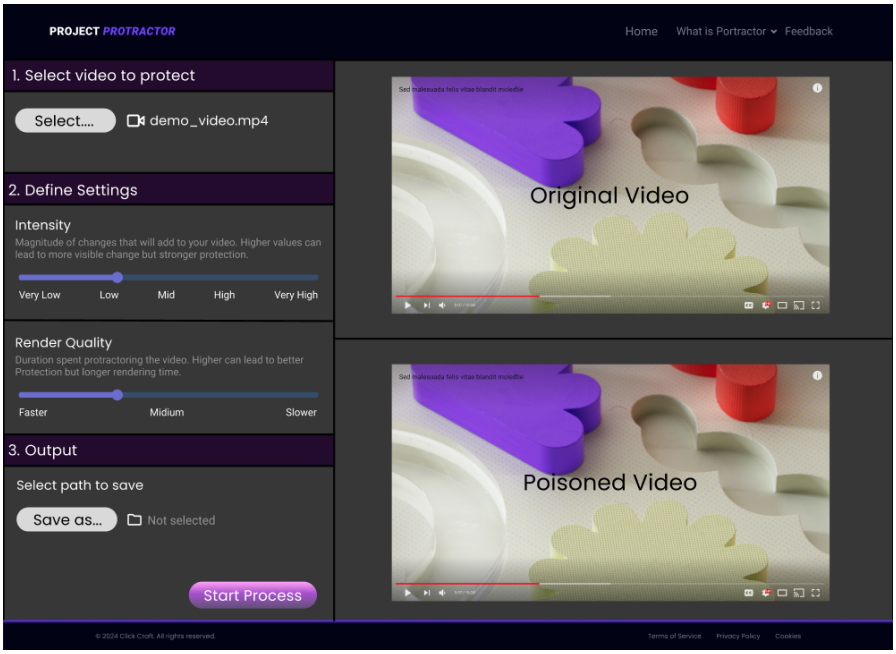

Figure 3.4: Mockup of Main Process Design

Chapter 4

Software Architecture Design

<TIP: Describe how you design your application using Unified Modelling Language (UML). There should be at least two diagrams that describe the software architecture. You may add additional or remove unnecessary diagrams. However, there needs to be a coherency between them at the end./>

4.1 Domain Model

<TIP: Describe the business concept of your project. Showcase a domain model that captures the said concept./>

4.2 Design Class Diagram

<TIP: Showcase a design class diagram for your project and explain how it works here. You can group classes into packages or layers to communicate your design better./>

4.3 Sequence Diagram

<TIP: Sequence diagrams describe how the software runs at runtime. You do not have to create a sequence diagram for every scenario. However, there should be one for all the main ones./>

<ChatGPT: Creating a sequence diagram for every use case is not strictly necessary, but it can be a valuable tool in certain situations. Sequence diagrams are particularly useful for illustrating the interactions

between different components or objects in a system over time, showcasing the flow of messages or actions between them./>

4.4 AI Component old

4.4.1 Poisoning Component

1. Generate Universal Adversarial Perturbation with stAdv and BTC-UAP perturbation techniques
2. Combine Input and Generated perturbation and limit them to stay within the color value range.
3. Ensure Output Quality with LPIPS and SSIM perceptual similarity ratings that users can set perceptual difference threshold.
   (a) LPIPS : ranges from 0 to infinity. High perceptual similarity is close to 0.
   (b) SSIM : ranges from -1 to 1. High perceptual similarity is close to 1.

4.4.2 Hardware Optimization Component

Before the Poisoning Process begins, perform hardware checks to minimize processing time

1. detect hardware's available CPU, GPU, SSD
2. maximize inference distribution across available CPU / GPU by cutting Videos into smaller videos and queue through different inferences
3. use SSD to alleviate cache from AI
4. maximize available batch for GPU
5. Eliminate identical frames to pick only one nearest identical frame to process through AI Pipeline with marked timestamp

Chapter 5

AI Component

5.1 Business Context AI Integration

Objective:

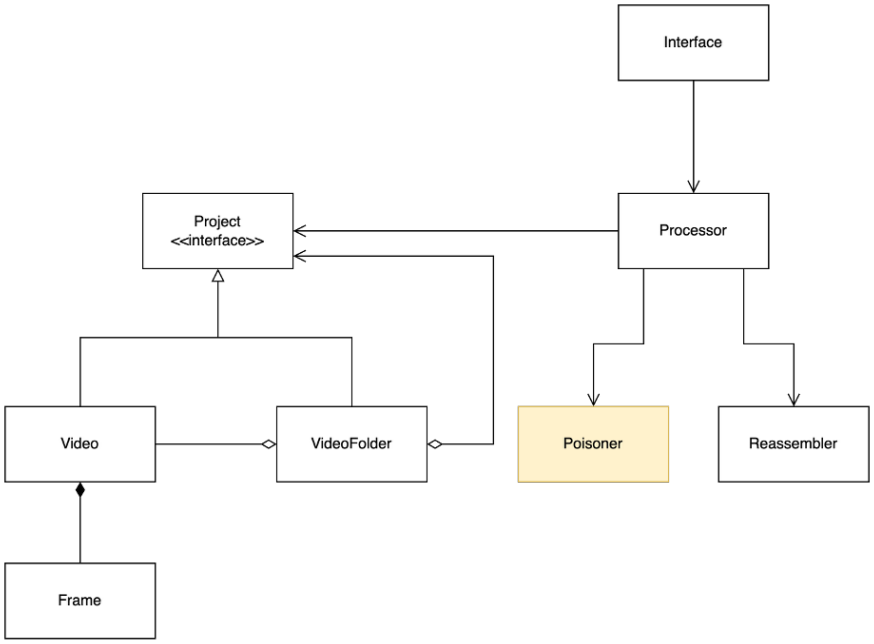

Figure 5.1: System's Domain Diagram

This is the system's domain diagram. The Poisoner component represents the part where AI will be used to generate and inject adversarial perturbations into the frames of input videos.

AI is suitable for this task because the 'poison' needs to be used on any video content, and remain effective as an adversarial attack against AI training while not being noticeably different from the original video. This requires the 'poison' to adapt to every possible input video's features

in light, color, etc., which is impractical to implement this feature with traditional programming alone due to the unpredictability and constantly changing nature of video content. By this way, we can ensure that the poisoning method will remain robust against AI training for every possible video.

5.2 Goal Hierarchy

5.2.1 Project Goals

Organization Goals

1. **Goal:** Protect creators' intellectual property from unauthorized AI training.
   Measured by: Reduction in cases of unauthorized AI usage reported by creators.
2. Establish the organization as a leader in AI-powered copyright defense.
   Measured by: Positive media mentions and industry recognition.
3. Encourage ethical AI practices and creator empowerment.
   Measured by: Number of creators adopting and actively using the system.

System Goals

- Process various video formats and inject adversarial perturbations with minimal perceptual change.
  Measured by: Percentage of processed videos meeting quality and poisoning effectiveness standards.
- Ensure poisoned videos degrade AI training effectiveness.
  Measured by: Testing AI model accuracy before and after poisoning, expecting a significant drop.
- Operate reliably and efficiently at scale.
  Measured by: System uptime and average processing time per video.

User Goals

- Enable users to easily upload, process, and download poisoned videos.
  Measured by: Number of successful video uploads and downloads.
- Provide users with confidence that their content remains visually unchanged while being protected.
  Measured by: User satisfaction ratings and survey feedback.
- Offer transparency and feedback on the poisoning process.
  Measured by: Number of support tickets or complaints related to output quality.

AI Model Goals

- Generate adaptive adversarial perturbations tailored to each video's unique features (light, color, texture, etc.).
  Measured by: Success rate of adversarial perturbations against various AI models.
- Balance invisibility of perturbation with high poisoning effectiveness.
  Measured by: PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) to ensure minimal perceptual difference.
- Maintain robustness against AI adaptation or retraining attempts.
  Measured by: Resistance testing against new AI models retrained on poisoned datasets.

5.3 Task Requirements Analysis Using AI Canvas

5.3.1 AI Task Requirements

Requirements (REQ)

- REQ-1: The AI must generate adversarial examples from input images using the FGSM technique.
- REQ-2: The AI should ensure poisoned images cause significant misclassification in target neural network models.

- REQ-3: The poisoning should preserve visual integrity such that it remains perceptually similar to the original.
- REQ-4: The AI must support batch processing for datasets.

Specifications (SPEC)

- SPEC-1: The AI uses the Adversarial Robustness Toolbox (ART) to implement FGSM attacks.
- SPEC-2: The perturbation strength ε should be configurable by the user.
- SPEC-3: Poisoning will be applied directly to image tensors prior to model training.

Environment (ENV)

- ENV-1: The system runs in a Python environment (e.g., Jupyter or Colab) with TensorFlow or PyTorch backends.
- ENV-2: Input data consists of image files (.jpg, .png)
- ENV-3: Requires GPU acceleration for efficient batch poisoning.
- ENV-4: May be extended later for use in video frame sequences.

5.3.2 AI Canvas Development

5.4 User Experience Design with AI

5.4.1 Interaction Style

Interaction Style: Annotate

- The AI modifies and tags poisoned images with visual indicators (e.g., warning labels).
- Users are informed when images are altered for adversarial purposes.
- This method provides transparency and keeps the user in control.

Canvas Element	Description
Input Data	Raw image files. Input will be uploaded manually.
Expected Output	Poisoned images with imperceptible perturbations that cause target models miss understand.
Tools/Frameworks	Python, ART, TensorFlow or PyTorch, NumPy, Google Colab or local machine with GPU.
Success Criteria	• Target models miss understand the feature. • No noticeable visual difference in the poisoned training images (PSNR > 30dB).

5.4.2 User Feedback Mechanism

Users are encouraged to provide feedback to help improve the AI poisoning system. The application offers a dedicated **Feedback** dropdown menu with the following options:

- **Comments and Suggestions** – Share thoughts or ideas for improving the poisoning algorithm.
- **Bug Reports** – Report technical issues or incorrect behavior in the poisoning process.
- **Email Us** – Directly contact the development team for in-depth support.

5.4.3 AI Contribution to System Intelligence

This system integrates AI to perform intelligent poisoning of image data. This AI component significantly enhances the system's capabilities beyond what is possible with non-AI approaches.

Without AI Integration

- Manual image editing to distort data.

- Watermarking or fixed filters that are easily bypassed by neural networks.

With AI Integration

- Automated poisoning of images based on gradient information.
- Adaptability to different neural network architectures.
- Preservation of visual quality for human viewers while deceiving models.
- Scalable batch processing of datasets with consistent adversarial effectiveness.

Chapter 6

Software Development

6.1 Software Development Methodology

<TIP: Describe your software development methodology in this section. /> Extreme Programming (XP) In short, it is - -Extreme programming (XP) is an Agile project management methodology that targets speed and simplicity with short development cycles. XP uses five guiding values, five rules, and 12 practices for programming. The structure is rigid, but the result of these highly focused sprints and continuous integrations can result in a much higher quality product.

- asana's summary of XP

By definition, XP is ...

In this project's case,

has 3 professor to overview work consistently 2 people project fixed deadlines 4 months project work time in total consistent testing single codebase

6.2 Technology Stack

<TIP: Describe your technology stack here. See the following example from ThaiProgrammer.org />

6.3 Coding Standards

<TIP: Describe your coding standard for this project here. />

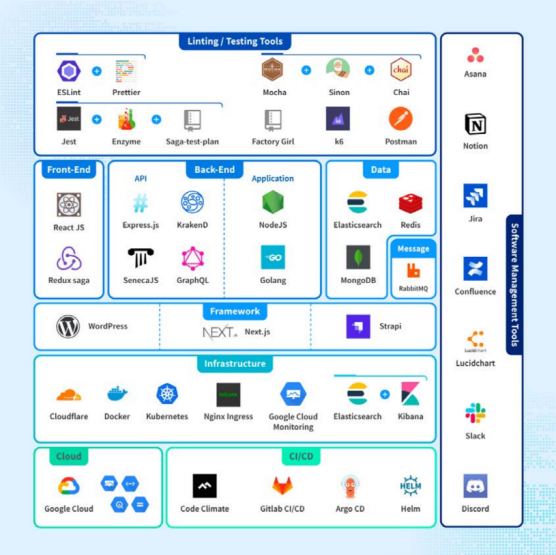

Figure 6.1: Example technology stack

6.4 Progress Tracking Report

<TIP: Show that you have been working on this project overtime. It can be in the form of a burndown chart or a contribution graph from GitHub./>

Chapter 7

Deliverable

7.1 Software Solution

<TIP: Share a link to your Github repository. Showcase screenshots of the application and briefly describe each page here. />

7.2 Test Report

<TIP: Describe how you test your project. Place a test report here. If you use continuousintegration and deployment (CI/CD) tools, describe your CI/CD method here. />

Chapter 8

Conclusion and Discussion

<TIP: Discuss your work here. For example, you can discuss software patterns that you use in this project, software libraries, difficulties encountered during development, or any other topic. />

We both need to develop a better base understanding of our AI generation, Data poisoning and many other topics to understand how we could poison video best, and as hardware efficient as possible. Thus, we currently are doing AI workshop labs in our freetime, advised by our project overseer Dr. Punpiti, starting since 7 January 2025.

AI labs - Cat-dog Classification - stAdv application - text-to-video training - BTC-UAP poisoning result and statistics - Quantifiable Metrics of video generation model degradation via adversarial attack

We have to be clear in privacy and use of customer data, where we confirm at no point in the development of this project we infringed existing copyright or user's data. User's data will not be collected to use nor bypass their own copyright. Any use will be towards to display degradation of poisoned model but the data itself will not be recorded nor made public to any identity. In any case, user can choose to opt out in the settings regarding the privacy of their contents, which might disabled some UI features.

We are new to many tech such as Electron js Huggingface and NVIDIA NIM, along with pytorch <queue system in python lib>

structure design method overall approved by advisors

Reference

Bibliography

[1] Overleaf, “Learn latex in 30 minutes,” https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes.

Appendix A

Appendix A: Example

<TIP: Put additional or supplementary information/data/figures in appendices. />

Appendix B

Appendix B: About LaTeX

LaTeX (stylized as LaTeX) is a software system for typesetting documents. LaTeX markup describes the content and layout of the document, as opposed to the formatted text found in WYSIWYG word processors like Google Docs, LibreOffice Writer, and Microsoft Word. The writer uses markup tagging conventions to define the general structure of a document, to stylize text throughout a document (such as bold and italics), and to add citations and cross-references.

LaTeX is widely used in academia for the communication and publication of scientific documents and technical note-taking in many fields, owing partially to its support for complex mathematical notation. It also has a prominent role in the preparation and publication of books and articles that contain complex multilingual materials, such as Arabic and Greek.

Overleaf has also provided a 30-minute guide on how you can get started on using LaTeX. [1]