Dissertation on

"Detection and Mitigation of Multimodal Privacy Leaks in AI Systems and Social Platforms"

Submitted in partial fulfillment of the requirements for

Bachelor of Technology
in
Computer Science & Engineering

UE23CS320A – Capstone Project Phase - 1

Submitted by:

K L Sonika	**PES2UG23CS247**
Gowni Ananya	**PES2UG23CS204**
Pranav Gaonkar	**PES2UG23CS426**
Nitish G	**PES2UG23CS402**

Under the guidance of

Dr. Geetha Dayalan
Associate Professor
PES University

Aug-Dec 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

'Detection and Mitigation of Multimodal Privacy Leaks in AI Systems and Social Platforms'

is a bonafide work carried out by

K L Sonika	**PES2UG23CS247**
Gowni Ananya	**PES2UG23CS204**
Pranav Gaonkar	**PES2UG23CS426**
Nitish G	**PES2UG23CS402**

In partial fulfillment for the completion of Fifth-semester Capstone Project Phase - 1 (UE23CS320A) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Aug 2025 – Dec. 2025. It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 5th-semester academic requirements in respect of project work.

Dr. Geetha Dayalan Associate Professor	Dr. Sandesh B J Chairperson

External Viva

Name of the Examiners	**Signature with Date**
1. _____	_____
2. _____	_____

DECLARATION

We hereby declare that the Capstone Project Phase - 1 entitled **"Detection and Mitigation of Multimodal Privacy Leaks in AI Systems and Social Platforms"** has been carried out by us under the guidance of **Dr. Geetha Dayalan, Associate Professor** and submitted in partial fulfillment of the course requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester Aug. – Dec 2025. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES2UG23CS247	**K L Sonika**
PES2UG23CS204	**Gowni Ananya**
PES2UG23CS426	**Pranav Gaonkar**
PES2UG23CS402	**Nitish G**

ACKNOWLEDGEMENT

I would like to express my gratitude to **Dr. Geetha Dayalan,** Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE23CS320A - Capstone Project Phase – 1.

I am grateful to all Capstone Project Coordinators, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Sandesh B J, Professor & Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department. I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Late Dr. M. R. Doreswamy, Founder, PES University, whose vision and dedication continue to inspire generations of learners. I would also like to express my sincere gratitude to Prof. Jawahar Doreswamy, Chancellor, PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University and Prof. Nagarjuna Sadineni, Pro Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this project could not have been completed without the continual support and encouragement I have received from my family and friends.

ABSTRACT

The rapid growth of multimodal AI systems has transformed automation, content creation, and digital communication. Models that jointly process text, images, audio, and video have enhanced user interaction but have also opened new pathways for privacy violations. In many cases, users unintentionally reveal sensitive Personally Identifiable Information (PII) such as facial features, voice signatures, names, addresses, background location details, and on-screen text present in images or videos.

LeakWatch is proposed as a middleware framework designed to intercept and sanitize such multimodal content **before** it is processed by AI models or shared on online platforms. The framework integrates a GraphSAGE-based Graph Neural Network for cross-modal detection of sensitive entities, a GAN-driven module for content redaction and replacement, and explainability components that provide transparent reasoning and auditability.

Phase 1 of this project covers the foundational work: reviewing existing literature, identifying gaps in current privacy-preservation techniques, studying suitable datasets, designing the system architecture, and outlining the methodology. This report summarizes the motivation, current research landscape, dataset preparation, and the planned workflow for Phase 2.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Multimodal Artificial Intelligence has grown rapidly, allowing systems like GPT-4o, LLaVA, and Gemma-Vision to understand and generate content across text, images, audio, and video. These capabilities power modern applications such as digital assistants, media generation, smart monitoring, customer support, and navigation.

However, this progress also brings new privacy risks. Users often upload multimodal content that contains sensitive information without realizing it — for example:

- Faces and other biometric details
- Background elements revealing location
- Voice characteristics that identify a speaker
- Screenshots with phone numbers, emails, or addresses
- Videos combining multiple privacy cues across frames

Most existing AI tools do not have strong, built-in mechanisms to filter or sanitize such content before it reaches the model. As a result, the sensitive data may be exposed, stored, or even memorized by large models, making it vulnerable to extraction attacks or unintended leakage during downstream use.

LeakWatch is designed to solve this issue by introducing a dedicated privacy-protection layer that operates before any multimodal data is processed by AI systems or social platforms. This middleware analyzes the input, detects potential privacy risks, and mitigates them through controlled transformations. By doing so, LeakWatch supports safer AI deployment and helps ensure that multimodal content is handled responsibly, securely, and in accordance with privacy standards.

CHAPTER 2

PROBLEM STATEMENT

In today's AI ecosystem, multimodal data processing has expanded across social platforms, enterprise systems, and consumer-facing applications. Users commonly interact with AI systems by uploading images, short videos, voice clips, and screenshots. The problem arises when users unknowingly expose sensitive Personally Identifiable Information (PII) that may be processed, stored, or inferred by AI models without explicit consent.

Key contributors to this problem include:

1. Lack of a unified multimodal privacy detection mechanism.
2. Inadequate mitigation strategies, such as naive blurring or blocking.
3. Absence of explainability in existing privacy pipelines.
4. Increasing adversarial sophistication, with malicious actors embedding hidden text, modifying faces, and manipulating audio signals.
5. Limited dataset availability for multimodal privacy research.

The LeakWatch Framework aims to solve these challenges by introducing an intelligent middleware capable of high-quality detection, robust mitigation, and transparent auditing, ensuring user data is safe before reaching the target system.

CHAPTER 3

LITERATURE SURVEY

3.1 Introduction to Multimodal Privacy Research

Multimodal privacy has emerged as a critical research area due to rapid advances in Vision-Language Models (VLMs), Large Language Models (LLMs), and diffusion-based generative systems. Existing works have evaluated vulnerabilities in text, image, audio, and video modalities, revealing that modern AI systems often leak Personally Identifiable Information (PII) and lack robust mitigation mechanisms. This chapter presents a detailed literature survey of eight key research papers, evaluating their methodology, results, and limitations to identify gaps relevant to the development of the LeakWatch Framework.

3.2.1 "Defeating Cerberus: Concept-Guided Privacy-Leakage Mitigation in Multimodal Language Models" (Zhang et al., 2025) [1]

3.2.1.1 Methodology

The authors proposed Concept-Guided Mitigation (Steering) using PCA applied to VLM internal latent representations. This method enforces refusal behavior for sensitive PII such as email addresses, names, and location data. The approach operates in a zero-shot manner with no retraining required.

3.2.1.2 Results

The model achieved an average 93.3% refusal rate on PII-related tasks while maintaining utility on non-PII tasks.

3.2.1.3 Limitations

The approach is limited to refusal/blocking and does not address multimodal privacy mitigation such as image redaction, selective blurring, or generative inpainting.

3.2.2 "Shake to Leak: Fine-tuning Diffusion Models Can Amplify the Generative Privacy Risk" (Li et al., 2024) [2]

3.2.2.1 Methodology

Li et al. introduced the Shake-to-Leak (S2L) attack, demonstrating that fine-tuning diffusion models significantly amplifies memorization. Synthetic sensitive data was used to fine-tune models and evaluate extraction risk.

3.2.2.2 Results

The attack increased MIA success by 5.4% AUC and enabled reconstruction of 15.8 private samples, proving that fine-tuned diffusion models retain sensitive visual information.

3.2.2.3 Limitations

The study analyzes only diffusion-model attack vectors and provides no defense or mitigation, highlighting a gap in protective preprocessing (the goal of LeakWatch).

3.2.3 "User Inference Attacks on Large Language Models" (Kandpal et al., 2024) [3]

3.2.3.1 Methodology

The authors used a User Inference Threat Model and Likelihood Ratio Test Statistic (LRTS) to determine if a user's data was present in an LLM's fine-tuning dataset.

3.2.3.2 Results

Experiments showed AUROC 88% attack success, revealing that fine-tuned LLMs overfit user-specific patterns that can be extracted.

3.2.3.3 Limitations

The work is restricted to text-only LLMs and does not consider multimodal leakage or visual/audio vulnerabilities.

3.2.4 "PROGAP: Progressive Graph Neural Networks with Differential Privacy Guarantees" (Sajadmanesh & Gatica-Perez, 2023) [4]

3.2.4.1 Methodology

PROGAP adopts Progressive Training and Aggregation Perturbation (AP) to achieve differential privacy in GNNs while conserving privacy budgets.

3.2.4.2 Results

The method achieved 5–10% higher accuracy than earlier DP-GNN baselines while maintaining privacy guarantees.

3.2.4.3 Limitations

Results apply only to graph-structured datasets, not multimodal data streams. It does not address runtime PII detection or mitigation.

3.2.5 "Assessing Visual Privacy Risks in Multimodal AI: A Taxonomy-Grounded Evaluation" (Tsaprazlis et al., 2025) [5]

3.2.5.1 Methodology

The study introduced a GDPR/CCPA-grounded Visual Privacy Taxonomy and evaluated VLMs on zero-shot privacy recognition tasks.

3.2.5.2 Results

Models demonstrated high inconsistency and poor sensitivity to privacy signals in images, exposing weaknesses in current AI systems.

3.2.5.3 Limitations

The work only evaluates privacy risk; it does not propose mitigation or preprocessing approaches.

3.2.6 "DRAG: Dynamic Region-Aware GCN for Privacy-Leaking Image Detection" (Yang et al., 2022) [6]

3.2.6.1 Methodology

The authors used Dynamic Region-Aware Graph Convolution (GCN) to detect privacy-leaking regions beyond object boundaries, including textures and background cues.

3.2.6.2 Results

DRAG achieved 87% accuracy, outperforming object-centric detectors.

3.2.6.3 Limitations

The model supports only image-based privacy detection and fails when external social or textual context is required.

3.2.7 "Unveiling Privacy Risks in Multimodal Large Language Models" (Chen et al., ACL 2025) [7]

3.2.7.1 Methodology

Chen et al. defined Disclosure and Retention Risks for multimodal LLMs and developed the MM-Privacy dataset evaluated across five task categories.

3.2.7.2 Results

Privacy leakage was found to be highly inconsistent across tasks; indirect tasks could bypass safety filters.

3.2.7.3 Limitations

The study is diagnostic and does not propose practical mitigation techniques.

3.2.8 "Advancing Content Moderation: Evaluating LLMs for Sensitive Content Detection Across Text, Images, and Videos" (AlDahoul et al., 2024) [8]

3.2.8.1 Methodology

The authors evaluated multimodal transformers (CLIP, ViLT, VideoCLIP) for sensitive-content detection using text-image-video fusion embeddings.

3.2.8.2 Results

Multimodal models performed better than single-modality models but struggled with context understanding and cross-modal consistency.

3.2.8.3 Limitations

The study focuses only on detection, not mitigation. It also lacks audio modality support and explainability.

3.3 Summary of Literature Insights

The reviewed literature clearly indicates that while numerous studies analyze sensitivity detection, inference attacks, and privacy vulnerabilities in multimodal AI, none provide a unified privacy-preserving framework. Most works lack mitigation, adversarial robustness, multimodal coverage, and interpretability—problems directly addressed by the proposed LeakWatch Framework, which integrates detection, mitigation, and explainability across text, images, audio, and videos.

CHAPTER 4

RESEARCH GAP

4.1 Introduction

Through a detailed examination of existing multimodal privacy research, multiple gaps were identified across detection, mitigation, explainability, dataset availability, and adversarial robustness. These gaps highlight the inadequacy of current systems when handling sensitive information embedded in text, images, audio, and video inputs. While modern AI models such as VLMs, LLMs, and diffusion systems have significantly advanced in capability, they still lack mechanisms that guarantee safety, privacy preservation, and regulatory compliance. The following subsections provide a comprehensive analysis of the critical research gaps revealed during the literature survey. [1]-[8]

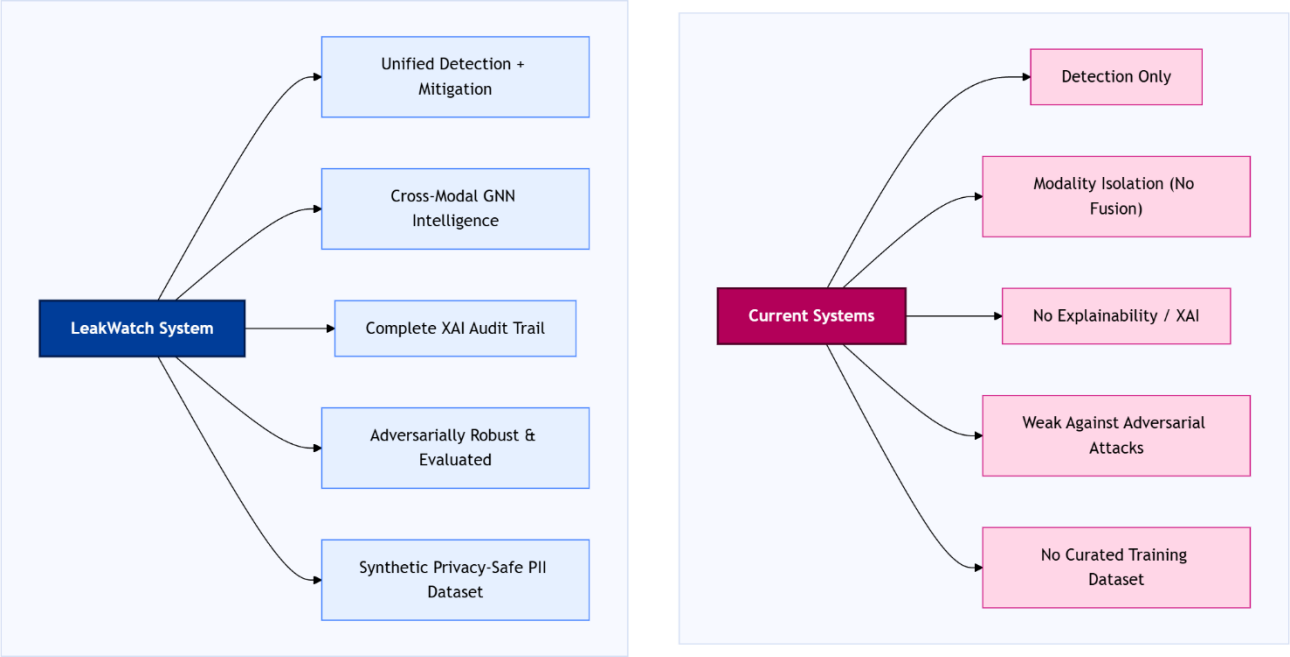

Figure 4.1: Research Gap Visualization

4.2 Gap 1 – Detection Without Mitigation

Most current studies focus heavily on the detection of privacy leaks, such as identifying faces, sensitive text, or audio signatures within multimodal content. While detection accuracy has improved, these methods stop short of addressing how sensitive content should be handled after identification. There is no unified mechanism for redacting, masking, or transforming the detected PII into a privacy-safe format. Techniques such as blurring, GAN-based synthesis, and inpainting have not been integrated into multimodal privacy systems [1], [2], [7], [8]. This creates a major vulnerability, as detection alone cannot prevent leakage—true privacy protection requires actionable mitigation steps, which current systems fail to provide.

4.3 Gap 2 – Modality Isolation

Existing privacy-preserving approaches generally treat each modality independently. For example, image-based systems focus on visual cues, while text-based models analyze language alone. However, privacy leaks in real-world scenarios are often multimodal in nature—such as a caption describing a person in an image, or an audio clip revealing location context that matches visual content. Current frameworks do not correlate multimodal information, leading to incomplete or inaccurate privacy assessment [3], [5], [6], [8]. A unified approach that links text, images, audio, and video entities is required to fully understand and mitigate privacy risks.

4.4 Gap 3 – Absence of Explainable Privacy Decisions

A significant deficiency across existing research is the lack of explainability. Users and regulatory bodies require clear justification for why specific elements were flagged as sensitive. Most AI models behave as black-box systems that provide outputs without human-understandable reasoning. Without interpretability tools such as Integrated Gradients, attention heatmaps, or subgraph extraction, users may lose trust in the system or misinterpret its decisions. Explainable privacy decisions are also

essential for compliance with GDPR, CCPA, and similar regulations, which require transparency and auditability—features missing in current multimodal frameworks [1], [5], [7].

4.5 Gap 4 – Weak Adversarial Defense

Multimodal systems are vulnerable to adversarial manipulation across multiple channels. Hidden-text attacks, steganography, adversarial image perturbations, synthetic voice cloning, and manipulated video frames can bypass traditional detection models. Studies reveal that even small adversarial changes can cause AI models to misinterpret content or overlook sensitive information. Existing frameworks focus on clean datasets and do not incorporate stress-testing against adversarial input [2], [6]. Without adversarial robustness, multimodal privacy systems remain fragile and unsuitable for deployment in security-sensitive environments.

4.6 Gap 5 – Dataset Limitations

There is a lack of large-scale, multimodal datasets explicitly designed for privacy research. Most benchmarks either focus on single-modality tasks or lack proper annotations for PII attributes. Without high-quality datasets that span text, images, audio, and video, it is difficult to train and evaluate multimodal privacy systems effectively. Synthetic augmentation, such as generating fake names, faces, ID cards, and adversarial audio signals, is required to overcome this limitation. However, current research does not provide standardized procedures for multimodal PII data generation [4], [7].

4.7 Conclusion

The identified research gaps emphasize the need for a unified, multimodal privacy-protection framework capable of detection, mitigation, explainability, and adversarial robustness. These insights form the foundation for the LeakWatch architecture, which aims to address all missing components in a single integrated pipeline.

CHAPTER 5

OBJECTIVES

5.1 Introduction

The core objectives of the LeakWatch Framework are formulated to directly address the research gaps outlined in Chapter 4. Each objective plays a key role in developing a reliable multimodal privacy-preserving middleware that can protect sensitive user information before it is accessed or processed by AI models or social online platforms.

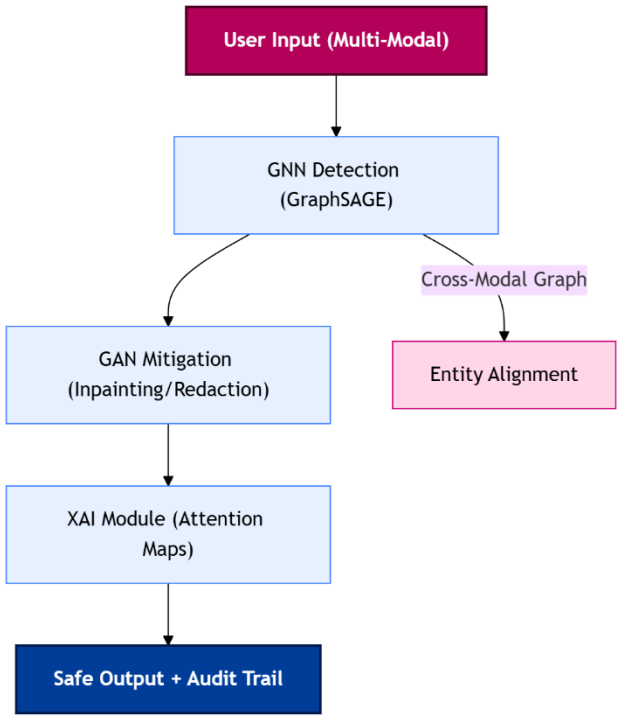

Figure 5.1: LeakWatch High-Level Architecture

5.2 Detailed Objectives

5.2.1 Multimodal Privacy Detection Architecture

To design a multimodal privacy detection architecture capable of correlating privacy-sensitive information across text, images, audio, and video. This involves feature extraction modules, entity alignment, and Graph Neural Network-based relationship modeling to identify complex privacy connections between modalities.

5.2.2 Graph Neural Network-Based Detection

To employ GraphSAGE and other GNN techniques for cross-modal entity detection, enabling understanding of contextual relationships such as linking a spoken name in audio to a face in an image or a location mentioned in text.

5.2.3 High-Quality Mitigation Techniques

To develop GAN-based and diffusion-based mitigation modules that perform selective redaction, inpainting, de-identification, and synthetic replacement of sensitive content. This ensures privacy without degrading user experience or utility.

5.2.4 Explainable AI Integration

To integrate Explainable AI tools such as Integrated Gradients, attention maps, and subgraph extraction to provide transparent, human-understandable explanations for every privacy decision made by LeakWatch.

5.2.5 Adversarial Robustness Development

To enhance adversarial robustness by stress-testing the system using GAN-generated attacks, hidden-text injections, modified faces, and adversarial audio signals. This ensures the system remains secure under real-world threats.

5.2.6 Dataset Preparation and Synthetic PII Augmentation

To prepare and augment multimodal datasets, including COCO, LibriSpeech, and Kinetics-400, with synthetic PII attributes. These augmentations enable training and evaluation of privacy-aware multimodal systems under diverse scenarios.

5.2.7 Regulatory Compliance and Safety

To ensure alignment with global privacy regulations such as GDPR and CCPA by incorporating transparent audit trails, user consent modeling, and data minimization principles.

5.2.8 Deployment-Ready Middleware Design

To develop a deployment-ready privacy-preserving middleware that integrates seamlessly with AI assistants, social media platforms, healthcare systems, and enterprise applications, providing a secure preprocessing layer.

CHAPTER 6

OVERVIEW OF DATASETS

6.1 Introduction

This chapter provides a detailed overview of the datasets used for developing and evaluating the LeakWatch multimodal privacy-preservation framework. Since privacy leaks can occur through **text, images, audio, and video**, it is essential for the dataset selection to represent all major modalities.

The chosen datasets focus on:

- **Real-world context**
- **Rich semantic content**
- **Potential PII exposure**
- **Suitability for synthetic augmentation**
- **Compatibility with GNN + GAN + Explainability models**

Because publicly available multimodal PII datasets are extremely rare, **Phase–1 incorporates synthetic PII augmentation** to simulate realistic privacy leaks.

The following sections describe each dataset in detail.

Dataset	Modality	Size	PII Potential	Augmentation Ready
Enron Emails	Text	600K+ emails	High (names, numbers, salary)	Yes (GAN text)
COCO	Image	330K images	High (faces, signs, docs)	Yes (overlay PII)
LibriSpeech	Audio	1000 hrs	Medium (spoken names)	Yes (voice cloning)
Kinetics-400	Video	240K clips	High (faces, screens)	Yes (frame overlay)

Table 6.1: Dataset Comparison

Modality	Before	After Mitigation (Concept)
Text	Hi John, my bank account number is 9876543210	Hi [PERSON], my bank account number is [BANK ACCOUNT NO.]
Image	[Face + License Plate]	[Blurred face + inpainted plate]
Audio	Waveform with "My credit card details are..."	Silence + beep
Video	Frame with ID card	Card region inpainted

Figure 6.2: Sample PII in Each Modality

6.2 Text Dataset – Enron Email Dataset

6.2.1 Description

The **Enron Email Dataset** is one of the largest publicly available corporate communication datasets, containing over **600,000 real emails** exchanged among Enron employees. It consists of genuine business conversations, forwarded threads, attachments, sender–receiver metadata, and semi-structured email bodies.

Enron is especially relevant for privacy-leak research because:

- Emails naturally contain **implicit and explicit PII**
- Sensitive corporate information appears in discussions
- Email threads create graph structures suitable for **GNN-based analysis**
- GANs can generate synthetic “dangerous PII leak” emails for augmentation
- The dataset is messy, real-world, and challenging — ideal for evaluating privacy leakage detection models

6.2.2 Data Attributes / Features

- 600K+ real corporate emails

- Sender–receiver communication graph
- Email threads (replies, forwards)
- Unstructured and noisy text (signatures, headers, informal language)
- Naturally occurring PII:
- Names
- Phone numbers
- Addresses
- Salary discussions
- Internal documents or confidential details
- Rich context for training and benchmarking text-based privacy detection models

6.3 Image Dataset - COCO

6.3.1 Description

The COCO (Common Objects in Context) dataset is a large-scale image dataset containing more than 330,000 images with detailed annotations. The dataset includes everyday scenes with humans, objects, environmental backgrounds, and contextual text such as signboards. This makes COCO valuable for studying privacy leaks in the visual modality, such as faces, location cues, personal belongings, and textual information embedded within images.

6.3.2 Data Attributes / Features

- High-resolution RGB images
- Detailed annotations (objects, masks, captions)
- People and identifiable facial features
- Background text (signboards, documents, labels)
- Complex multi-object scenes
- Useful for:

- Visual PII detection
- Face privacy analysis
- Text-in-image privacy leaks

6.4 Audio Dataset - LibriSpeech

6.4.1 Description

LibriSpeech is a corpus of approximately 1000 hours of English speech extracted from audiobooks. It contains diverse speaker identities, accents, pitch variations, and background noise patterns. For LeakWatch, LibriSpeech supports the identification of **audio-based PII leakage**, such as spoken names, addresses, phone numbers, and other sensitive verbal disclosures. It is also compatible with GAN-based synthetic audio augmentation.

6.4.2 Data Attributes / Features

- 16 kHz high-quality speech recordings
- Wide range of speaker identities
- Natural variations in pitch, rate, and prosody
- Narratives that may include extractable personal details
- Suitable for:
- Voice identity detection
- Spoken PII recognition
- Robust audio privacy models

6.5 Video Dataset - Kinetics-400

6.5.1 Description

Kinetics-400 is a large-scale video dataset consisting of short clips representing 400 different human actions. Each video clip contains movement, people, environments, and objects that may reveal PII

such as faces, surroundings, and social context. This dataset is crucial for studying privacy leaks in dynamic video content and for analyzing cross-modal dependencies between visual and motion cues.

6.5.2 Data Attributes / Features

- High-quality video clips (10 seconds each)
- Temporal motion information
- Multiple individuals per frame
- Realistic indoor & outdoor environments
- Natural presence of potential PII such as:
- faces
- documents
- digital screens
- ID cards/badges
- personal belongings
- Ideal for multimodal privacy-leak detection in video streams

6.6 Synthetic PII Augmentation

6.6.1 Description

Due to the lack of standardized multimodal PII datasets, Phase–1 includes a **synthetic augmentation pipeline** that introduces controlled privacy leaks across modalities. These augmentations allow the system to simulate:

- dangerous private information
- hidden leaks
- adversarial patterns
- complex multimodal interactions

GANs and automated overlays are used to inject realistic synthetic private information.

6.6.2 Data Attributes / Features

- Synthetic name, address, and number overlays on images
- Artificial identity cards, documents, bills, credit cards
- Mixed audio containing synthesized sensitive speech
- GAN-based face blending for anonymization
- Adversarial noise and perturbations
- Thread-level text augmentation for hidden PII in emails

These augmentations ensure robust training and evaluation of privacy-leak detection models across all modalities.

6.7 Conclusion

The selected datasets—Enron (text), COCO (image), LibriSpeech (audio), and Kinetics-400 (video)—provide a comprehensive and realistic foundation for multimodal privacy-leak detection. Combined with synthetic augmentation, they enable LeakWatch to handle diverse real-world scenarios involving hidden, explicit, and adversarial privacy risks.

Together, these datasets support the development of a robust, generalizable, and explainable multimodal privacy preservation framework.

CHAPTER 7

CONCLUSION OF CAPSTONE PROJECT PHASE 1

Phase 1 successfully laid the groundwork needed to develop the LeakWatch Framework. During this stage, the team completed the essential research survey, identified gaps in existing multimodal privacy solutions, finalized the problem definition, and selected appropriate datasets for text, image, audio, and video modalities. The conceptual architecture and methodological direction were also outlined. Together, these outcomes form a strong foundation for moving into Phase 2, where system design and model development will begin.

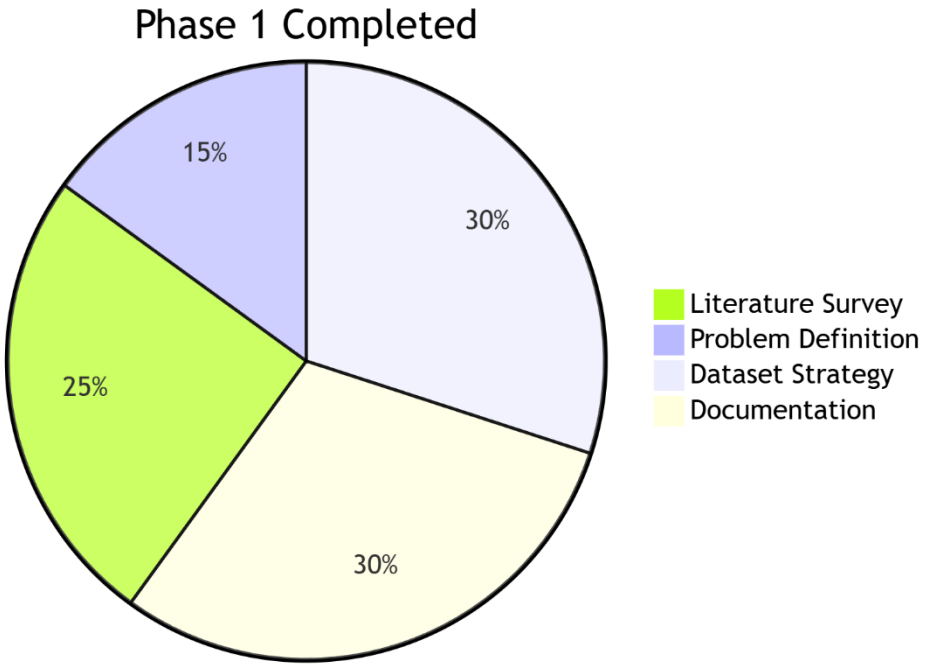

Figure 7.1: Phase 1 Achievements Dashboard

Main achievements are:

7.1 Literature Survey

During Phase 1, an extensive literature review was conducted across domains such as multimodal privacy leaks, adversarial attacks, GNN-based detection, GAN-based mitigation, and explainability methods.

The survey helped us:

- Understand how private information gets leaked through text, images, audio, and video.
- Analyze existing detection models and their limitations in real-world scenarios.
- Identify gaps such as inability to detect hidden/indirect PII, limited multimodal integration, and poor robustness against adversarial content.
- Define where our framework can contribute meaningfully, especially using GNN + GAN + Explainability.

This survey was the backbone of defining the direction of LeakWatch.

7.2 Final Problem Statement

Phase 1 concluded with a clear and academically strong problem statement:

"Detection and Mitigation of Multimodal Privacy Leaks in AI Systems and Social Platforms."

This statement captures the goal of LeakWatch—to identify sensitive information across multiple modalities and apply intelligent mitigation strategies without reducing content quality.
It sets the scope for text, image, audio, and video privacy protection, making the framework applicable to real-world social media, cloud platforms, and AI models.

7.3 Dataset Preparation

A structured dataset plan was created to support each modality of the framework:

- **Enron Email Dataset – Text**
  Contains real conversations, making it suitable for mining both explicit and hidden sensitive information.
- **COCO Dataset – Images**
  Large-scale images with people, objects, documents, and real-world scenes for detecting visual PII.
- **LibriSpeech – Audio**
  Human speech samples useful for identifying speaker-specific private information.
- **Kinetics-400 – Video**
  Video clips containing human activities, enabling frame-wise privacy leak studies.

In addition, Phase 1 introduced **synthetic adversarial augmentation** using GANs and LLM-based generation:

- Hidden text inside images
- Artificial ID cards, QR codes, documents
- Fake bank numbers or SSN-like patterns
- Mixed audio containing synthetic private information

This ensures the dataset becomes challenging and more suitable for advanced research.

7.4 Documentation

All Phase 1 activities were thoroughly documented, including:

- Literature survey summaries
- Problem statement justification
- Dataset collection strategies
- Synthetic augmentation design

- Observations, challenges, and planning notes

This documentation forms the official base for Phase 2 development and will support presentations, evaluations, and future publications.

7.5 Conclusion

Phase 1 successfully completed all groundwork required to build LeakWatch.

We now have:

- A strong research understanding
- A well-defined problem
- A multimodal dataset strategy
- Clear documentation

With these foundations in place, Phase 2 will focus on building the actual detection and mitigation pipeline using GNN, GAN, and Explainability modules.

CHAPTER 8

PLAN OF WORK FOR CAPSTONE PROJECT PHASE 2

Phase 2 focuses on developing the core functional modules of the LeakWatch Framework. The objective is to build the detection and mitigation pipeline, integrate the multimodal components, and prepare initial testing and documentation.

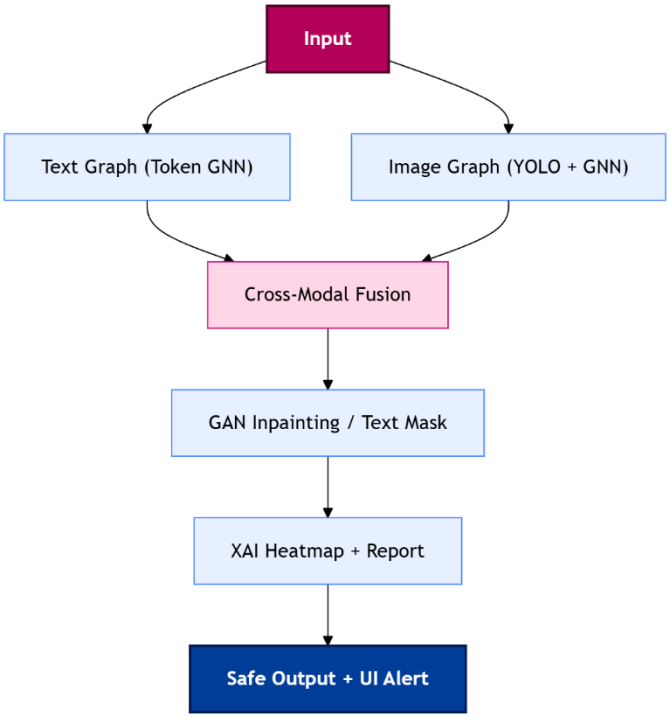

Figure 8.1: Phase 2 Pipeline Flow

8.1 High-Level System Design

In this stage, we define the complete architecture of LeakWatch for all modalities—text, images, audio, and video.

Key Activities:

- Create the full workflow diagram showing how data moves from input → detection → mitigation → safe output.
- Decide how the three major components will interact:
    - **GNN Module:** Detects sensitive entities using graph structure.
    - **GAN Module:** Modifies or removes detected sensitive regions.
    - **Explainability Module:** Highlights why the model flagged something as sensitive.

- Plan the communication layer between models (e.g., JSON results, bounding boxes, token indices).
- Finalize the data schemas for nodes, edges, token graphs, and image-object graphs.
- Ensure the framework remains extensible for audio and video in later stages.

8.2 GNN Detection for Text and Images

Phase 2 includes building the core detection engine using a GraphSAGE-based Graph Neural Network.

For Text:

- Convert each sentence/document into a **token graph**, connecting tokens based on semantic closeness, dependency relations, or positional adjacency.
- Label sensitive tokens such as:
    - Names, bank account numbers, PAN/Aadhar-like patterns, SSN-like patterns, locations, etc.
- Train and evaluate the GNN on:
    - Existing datasets (Enron Email Corpus)
    - synthetic PII augmentation (bank numbers, IDs, passwords, etc.).

For Images:

- Detect objects using Faster-RCNN or YOLO and treat each region as a **graph node**.
- Build connections based on object proximity or scene relationships.

- Train the GNN to classify sensitive regions:
    - Faces
    - ID cards
    - Credit cards
    - Screens/printed documents
    - License plates

8.3 GAN Mitigation for Text and Images

Once sensitive information is detected, the GAN module performs the actual privacy protection.

Text Mitigation

- Replace or mask sensitive tokens with:
    - Placeholder tags (e.g., <MASK_ID>)
    - Synthetic safe alternatives (fake names, fake account numbers)
- Ensure the modified text remains readable and natural.

Image Mitigation

- Use **Stable Diffusion inpainting** or a GAN-based inpainting model to remove/reconstruct sensitive regions.
- Target regions include:
    - Faces
    - ID cards
    - Documents
    - Number plates
- Ensure that the edited image looks visually consistent and non-distorted.

8.4 Basic Pipeline Testing

Before integrating everything, Phase 2 will test each modality separately.

Text Pipeline Testing

1. Input text →
2. GNN detects sensitive tokens →
3. GAN/Text Rewriter replaces or removes them →
4. Output clean text.

Image Pipeline Testing

1. Input image →
2. GNN detects sensitive bounding boxes →
3. Stable Diffusion inpaints/blur →
4. Output clean image.

Audio + Video (Initial Planning Only)

- Use placeholder modules to simulate:
  - Audio: simple keyword spotting for private info.
  - Video: frame-level image processing.
- No full integration in Phase 2; only proof-of-concept.

8.5 Documentation

Throughout Phase 2, documentation will be continuously maintained.

What to document:

- Architecture diagrams and module descriptions.
- Dataset details and augmentation steps.
- GNN model design, training logs, and accuracy metrics.
- Screenshots of image inpainting and text masking.

- Pipeline testing results.
- Challenges faced and solutions applied.
- Slides and report for the Phase 2 evaluation.

8.6 Conclusion

Phase 2 outlines the planned technical development for the LeakWatch framework. In this phase, we have defined what needs to be built next, including the GNN detection module, GAN-based mitigation module, and the complete multimodal processing pipeline. These components are not implemented yet, but their design, workflow, and integration steps have been clearly mapped out.

Overall, Phase 2 serves as a forward-looking plan that prepares the groundwork for actual implementation in the upcoming phase. It ensures that all modules, data flows, and testing procedures are well-structured and ready to be developed in Phase 3.

REFERENCES/BIBLIOGRAPHY

[1] B. Zhang, I. E. Akkus, R. Chen, A. Dethise, K. Satzke, I. Rimac, and Y. Zhang, "Defeating Cerberus: Concept-Guided Privacy-Leakage Mitigation in Multimodal Language Models," arXiv preprint arXiv:2509.25525, 2025. [Online]. Available: https://arxiv.org/abs/2509.25525

[2] Z. Li, J. Hong, B. Li, and Z. Wang, "Shake to Leak: Fine-tuning Diffusion Models Can Amplify the Generative Privacy Risk," arXiv preprint arXiv:2403.09450, 2024. [Online]. Available: https://arxiv.org/abs/2403.09450

[3] N. Kandpal, K. Pillutla, P. Kairouz, C. A. Choquette-Choo, A. Oprea, and Z. Xu, "User Inference Attacks on Large Language Models," arXiv preprint arXiv:2310.09266, 2024. [Online]. Available: https://arxiv.org/abs/2310.09266

[4] S. Sajadmanesh and D. Gatica-Perez, "PROGAP: Progressive Graph Neural Networks with Differential Privacy Guarantees," arXiv preprint arXiv:2304.08928, 2023. [Online]. Available: https://arxiv.org/abs/2304.08928

[5] E. Tsaprazlis, T. Feng, A. Ramakrishna, R. Gupta, and S. Narayanan, "Assessing Visual Privacy Risks in Multimodal AI: A Novel Taxonomy-Grounded Evaluation of Vision-Language Models," arXiv preprint arXiv:2509.23827, 2025. [Online]. Available: https://arxiv.org/abs/2509.23827

[6] G. Yang, J. Cao, Q. Sheng, P. Qi, X. Li, and J. Li, "DRAG: Dynamic Region-Aware GCN for Privacy-Leaking Image Detection," arXiv preprint arXiv:2203.09121, 2022. [Online]. Available: https://arxiv.org/abs/2203.09121

[7] T. Chen, P. Li, K. Zhou, T. Chen, and H. Wei, "Unveiling Privacy Risks in Multi-modal Large Language Models: Task-specific Vulnerabilities and Mitigation Challenges," in Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 4573–4586.

[8] N. AlDahoul, M. J. T. Tan, H. R. Kasireddy, and Y. Zaki, "Advancing Content Moderation: Evaluating Large Language Models for Detecting Sensitive Content Across Text, Images, and Videos," arXiv preprint arXiv:2411.17123, 2024. [Online]. Available: https://arxiv.org/abs/2411.17123

APPENDIX A

DEFINITIONS, ACRONYMS, AND ABBREVIATIONS**

A.1 Definitions

Multimodal Data

Data coming from multiple sources such as text, images, audio, and video.

Privacy Leak

Sensitive information that may reveal a person's identity (such as name, face, or phone number) that is accidentally exposed.

Middleware

A software layer placed between the user's input and the target AI system to filter, sanitize, or process data before forwarding it.

Inpainting

A technique used by image models to fill or replace parts of an image, typically to hide sensitive regions like faces or license plates.

Redaction

The process of removing or masking sensitive text such as names, phone numbers, or addresses.

Adversarial Input

Deceptive or manipulated content—such as hidden text in an image—intended to trick or bypass machine learning systems.

Audit Trail

A visual or textual explanation showing why and where the system detected sensitive information.

A.2 Acronyms and Abbreviations

Term	Full Form	Meaning in Project
AI	Artificial Intelligence	Systems that process multimodal data.
LLM	Large Language Model	Models like GPT-4o that consume multimodal inputs.
VLM	Vision-Language Model	Models that combine image and text understanding.
PII	Personally Identifiable Information	Sensitive data such as names, faces, phone numbers.
GNN	Graph Neural Network	Used for cross-modal privacy detection in LeakWatch.
GraphSAGE	Graph Sample and Aggregate	The specific GNN architecture used in the project.
GAN	Generative Adversarial Network	Used for mitigation and generating adversarial samples.
XAI	Explainable Artificial Intelligence	Provides explanations and audit trails.
DP	Differential Privacy	Technique to protect training data (from literature).
CLIP	Contrastive Language–Image Pretraining	Used for text-image feature extraction.
YOLO	You Only Look Once	Used for image object detection (faces, plates).
ASR	Automatic Speech Recognition	Converts audio to text (e.g., Whisper).
SSIM	Structural Similarity Index Measure	Metric to evaluate inpainted image quality.
BLEU	Bilingual Evaluation Understudy	Metric to evaluate text modification quality.
UI	User Interface	Social media confirmation screen for users.
API	Application Programming Interface	Used for connecting middleware to other systems.

Table A.2: Acronyms and Abbreviations with meaning in the project