Privacy-Preserving Inference for Vision-Transformers/Visual Language Models

An M. Tech Thesis Project report submitted in fulfillment of the Requirements for the Degree of

Master of Technology in Data Science

by

Avadhesh Sisodiya
(244161012)

under the guidance of

Prof. Gaurav Trivedi

&

Asst. Prof. Ayon Borthakur

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

Acknowledgement

I would like to express my sincere gratitude to my supervisors, **Prof. Gaurav Trivedi** and **Asst. Prof. Ayon Borthakur**, for their continuous guidance and support throughout this project, **Privacy-Preserving Inference for Vision-Transformers or Visual Language Models**. Their insightful suggestions, constructive feedback, and constant encouragement played a crucial role in helping me understand the subject and complete this work successfully. I am deeply thankful for the time and effort they invested in mentoring me during the course of this research.

Avadhesh Sisodiya
Roll No.: 244161012
M.Tech in Data Science
Indian Institute of Technology Guwahati
Guwahati, Assam 781039, India

Abstract

The increasing deployment of Vision Transformers (ViTs) in real-world applications has amplified concerns over data privacy, particularly when models process sensitive visual information such as identity documents or biometric attributes.

Vision Transformers (ViTs) achieve state-of-the-art performance on visual tasks but remain incompatible with secure **Multi-Party Computation (MPC)** due to their reliance on non-polynomial operations such as Softmax, GeLU, and LayerNorm. These functions are expensive to evaluate under MPC, making **privacy-preserving inference** impractical. In this work, we adopt the PriViT architecture, which selectively replaces nonlinearities using gradient-guided **Taylor polynomial approximations**, enabling fully MPC-friendly ViTs while maintaining high accuracy.

In Phase I, we reproduce and fine-tune the **PriViT ViT-Tiny model on CIFAR-10 and CIFAR-100, achieving 97.79% accuracy on CIFAR-10** with negligible degradation compared to the original transformer. These results demonstrate that Taylorized ViTs can preserve utility while becoming compatible with MPC protocols.

Building upon these findings, the next phase of our work aims to integrate MPC-friendly Vision Transformers into Visual Language Models (VLMs) to achieve **privacy-preserving multimodal inference**. Inspired by **PrivTune**, we plan to explore minimal-data fine-tuning strategies to enhance privacy awareness in VLMs. The long-term objective is to develop privacy-aligned VLMs for **Indian data ecosystems**, particularly for tasks involving sensitive personal information (e.g., passports, Aadhaar cards, and driving licenses), while ensuring compliance with emerging Indian privacy and data protection regulations.

Through this work, we aim to bridge the gap between secure computation and vision-language understanding, paving the way for high-utility, MPC-compatible AI systems capable of performing accurate yet confidential visual reasoning.

Contents

List of Tables

Chapter 1

Introduction

1.1 Introduction

The rapid proliferation of deep learning systems in high-stakes applications has intensified concerns around data privacy, security, and responsible AI deployment. Modern computer vision and multimodal architectures—particularly Vision Transformers (ViTs) and Visual Language Models (VLMs)—are increasingly being used for tasks involving sensitive personal information such as identification documents, biometric imagery, and confidential records. However, as these models grow in capability, they also pose heightened risks of data leakage and misuse if not handled within rigorous privacy-preserving frameworks.

Among privacy-preserving computation techniques, Secure Multi-Party Computation (MPC) has emerged as a strong candidate for enabling inference on encrypted or secret-shared data without exposing the underlying inputs. MPC distributes data among several computing parties in an encrypted or masked form and ensures that no individual party can reconstruct the original input while still permitting collaborative computation. This paradigm provides cryptographic privacy even against powerful adversaries, making it suitable for handling sensitive visual and multimodal datasets. However, one of the central problems in deploying modern deep learning models under MPC is that these architectures rely heavily on non-polynomial nonlinear operations—such as exponentials, error functions, square roots, and divisions—which translate poorly into the arithmetic circuits supported by MPC protocols. As a result, mainstream architectures like Vision Transformers become prohibitively expensive under MPC, limiting their usability in real-world secure inference settings.

To address this challenge, recent work such as PriViT (Privacy-Preserving Vision Transformer) proposes a solution that selectively replaces MPC-incompatible nonlinearities with polynomial approximations, enabling ViTs to operate using only additions and multiplications, which are efficient in MPC. PriViT introduces a gradient-based algorithm that identifies and Taylorizes nonlinear operations—specifically GELU activations, Softmax functions, and LayerNorm components—while carefully maintaining accuracy. By introducing learnable coefficients to control the degree of Taylorization at each layer, PriViT achieves a balance between privacy-compatibility and predictive performance. The key innovation lies in allowing the model to decide which nonlinear components can be approximated without significantly harming accuracy. This selective Taylorization significantly reduces MPC evaluation latency and enables Vision Transformers to become practically deployable for secure inference.

1.1.1 Motivation

The motivation behind this thesis emerges from the intersection of three domains: (1) the widespread use of Vision Transformers and VLMs, (2) the privacy risks posed by sensitive visual and multimodal data, and (3) the computational constraints of cryptographic protocols like MPC. While Convolutional Neural Networks (CNNs) have been traditionally favored for privacy-preserving inference due to their simpler computational graphs, they often underperform compared to transformers on complex tasks. Moreover, prior MPC-compatible models either significantly reduce model expressiveness or incur large accuracy drops. PriViT addresses these limitations by preserving the transformer's expressive capacity while making it MPC-compatible through polynomial approximation.

In the Indian context, the relevance of privacy-preserving AI has been amplified by the emergence of national datasets and large-scale digital identity systems. Documents such as Indian passports, Aadhaar cards, PAN cards, and other government-issued IDs contain sensitive demographic and biometric information. With the emergence of stronger data protection frameworks, such as the Digital Personal Data Protection (DPDP) Act, and evolving national privacy guidelines, there is a growing demand for AI systems that can operate securely on sensitive information without exposing or storing raw data.

Our goal is to build on PriViT and create MPC-compatible Vision Transformers, and eventually extend them to MPC-compatible Visual Language Models (VLMs).

This enables secure multimodal reasoning—for example, verifying document authenticity, extracting structured information from passports, or assisting digital governance processes—without exposing raw data to the model or service provider.

1.1.2 Research Problem

Despite the promise of MPC, integrating modern transformers into privacy-preserving pipelines remains challenging due to:

1. **Non-polynomial nonlinearities in ViTs**: Functions such as Softmax, GELU, and LayerNorm require expensive operations (exponentials, divisions, square roots), which are inefficient in MPC.
2. **High communication and computation costs**: MPC inherently requires multiple communication rounds. When each nonlinearity requires complex secure computation, inference latency becomes impractical.
3. **Accuracy degradation from naive approximation**: Replacing nonlinearities with simple polynomials often degrades accuracy. Achieving a balance between MPC-friendliness and utility requires careful optimization.
4. **Extending MPC-friendly architectures to VLMs**: VLMs combine vision encoders with large language models. Making them MPC-compatible requires lightweight tuning strategies, as full fine-tuning is computationally expensive.

The goal of this thesis is to solve these challenges by using the PriViT architecture as the foundation for developing efficient, accurate, and MPC-friendly transformer models.

1.1.3 Objectives

This thesis is divided into two phases:

Phase I: MPC-Friendly Vision Transformers

- Reproduce and implement the PriViT architecture using ViT-Tiny.

- Implement gradient-guided Taylorization of GELU and Softmax.
- Train and evaluate PriViT on CIFAR-10 and CIFAR-100 to verify accuracy retention.
- Analyze the nonlinearity counts (GELU and Softmax operations) before and after Taylorization.
- Evaluate model accuracy and MPC suitability.

Phase II: Extending to Secure Visual Language Models

- Explore minimal-tuning approaches inspired by PrivTune for privacy-aligned VLMs.
- Design an MPC-compatible VLM architecture using Taylorized ViT encoders.
- Apply the pipeline to sensitive Indian datasets (passport, Aadhaar, ID documents).
- Ensure compliance with Indian privacy and data protection laws.

1.1.4 Significance of the Work

This research carries both technical and societal importance:

- **Technical**: Enables high-performing transformer models to operate under MPC constraints; provides empirical insights into Taylorized VLMs.
- **National relevance**: Supports secure processing of sensitive Indian identification documents under emerging privacy laws.
- **Practical impact**: Facilitates confidential AI services in identity verification, finance, healthcare, and e-governance without exposing user data.

1.1.5 Thesis Organization

This thesis is structured as follows:

- **Chapter 1**: Introduction
- **Chapter 2**: Motivation and Problem Statement
- **Chapter 3**: Review of Prior Works
- **Chapter 4**: Literature Review
- **Chapter 5**: Methodology and Preliminary work
- **Chapter 6**: Conclusion and future work

Chapter 2

Motivation and Problem Statement

2.0.1 Motivation

The shift toward cloud-based AI services has led to the widespread adoption of client–server architectures, where computationally expensive models run on a remote server while the client transmits inputs for inference. However, this setup creates a fundamental privacy conflict: the *client* does not want to share confidential or personally identifiable data with an untrusted server, and the *server* does not want to expose its proprietary model parameters or allow the client to query the model in a way that enables model extraction. This mutual distrust motivates the need for *secure private inference*, where the client learns only the prediction and the server learns nothing about the client's input.

Traditional deep learning models for private inference have largely relied on CNN-based architectures due to their simple nonlinearities (e.g., ReLU), which are relatively easy to approximate in secure computation frameworks. However, modern vision systems have transitioned toward *Vision Transformers (ViTs)* and *Visual Language Models (VLMs)*, which now dominate state-of-the-art performance across image classification, image–text retrieval, document understanding, and multimodal reasoning tasks. Unfortunately, ViTs introduce a number of nonlinear operations—such as Softmax, GELU, LayerNorm, and exponential/erf-based functions—that are prohibitively expensive under Secure Multi-Party Computation (MPC). These operations require non-polynomial functions that translate poorly to arithmetic circuits and substantially increase communication overhead, making standard ViTs incompatible with practical MPC execution.

The PriViT architecture proposes an elegant solution to this challenge by *selectively Taylorizing* nonlinearities within a pretrained ViT model. Using gradient-guided masks (α and β parameters), PriViT learns which nonlinear components can be replaced with low-degree polynomial approximations while preserving accuracy. This approach maintains the original ViT architecture, reduces MPC latency, and avoids the need for expensive neural architecture search. PriViT therefore provides an attractive foundation for building *MPC-friendly Vision Transformers* suitable for privacy-preserving inference on sensitive visual data.

In addition to secure computation concerns, recent analyses of modern Visual Language Models highlight a complementary privacy problem: even when the input remains encrypted through MPC, the model itself may not understand or properly handle privacy-sensitive content. Studies show that current VLMs exhibit inconsistent behavior when confronted with faces, identity documents, or other private visual attributes. New benchmarks such as *PRIVBENCH* and *PRIVBENCH-H* demonstrate a lack of robust privacy awareness, whereas the *PRIVTUNE* dataset has shown that small amounts of carefully designed instruction tuning can greatly improve privacy sensitivity. This motivates our long-term goal of not only enabling MPC-friendly ViTs but also developing *privacy-aware VLMs* capable of safe and confidential multimodal reasoning over sensitive Indian datasets such as passports, Aadhaar cards, and financial documents.

2.0.2 Problem Statement

Although Secure Multi-Party Computation offers strong cryptographic guarantees for private inference, modern transformer-based architectures remain impractical to deploy under MPC due to their reliance on nonlinear operations that are computationally expensive to evaluate securely. The Softmax function in self-attention, the GELU activation in MLP layers, and the division and square root operations in LayerNorm all require non-polynomial computations that lead to excessive communication overhead and latency in MPC protocols. As a result, there is a pressing need to redesign or adapt Vision Transformers so that they can be executed efficiently under MPC without sacrificing accuracy.

Furthermore, as VLMs become integral to real-world applications, there is a growing concern that these models do not adequately recognize or handle privacy-sensitive content, even when encryption techniques protect the raw inputs. This dual challenge—MPC incompatibility of transformers and weak privacy awareness in current VLMs—motivates the need for architectures and training methodologies that jointly address secure computation and privacy-aware behavior.

Formal Problem Statement (in simple terms): Given a pretrained Vision Transformer model $f(x)$, the goal is to construct a new model $f'(x)$ such that:

1. $f'(x)$ can be executed using only MPC-friendly operations (additions and multiplications), by replacing nonlinearities with polynomial approximations where appropriate;
2. $f'(x)$ preserves the original predictive accuracy of $f(x)$ as much as possible;
3. the system can be extended toward VLMs that exhibit improved privacy-aware behavior when processing sensitive multimodal data;
4. neither the client learns the server's model, nor the server learns the client's input.

In essence, the problem is to build a Vision Transformer (and later a VLM) that can run securely under MPC, maintain high accuracy, and understand how to safely handle privacy-sensitive content.

Chapter 3

Review of Prior Works

Private inference for deep learning models has been extensively studied in the context of Convolutional Neural Networks (CNNs), but extending similar techniques to Vision Transformers (ViTs) introduces new challenges due to the presence of complex nonlinear operations such as Softmax, GELU, LayerNorm, exponentials, and square-root based computations. These operations are expensive to evaluate under Secure Multi-Party Computation (MPC), limiting the deployment of standard ViT architectures in privacy-preserving inference settings.

In recent years, several approaches have attempted to address this problem by modifying or approximating nonlinear components of ViTs to make them compatible with MPC. This chapter provides a detailed review of the three main published approaches: *MPCViT*, *SAL-ViT*, and *RNA-ViT*. We also discuss the *PriViT* framework, which forms the foundation of the work presented in this thesis.

3.1 Overview of Existing Approaches

Table 3.1 summarizes the key design elements of prior attempts to construct MPC-friendly Vision Transformers. These works differ significantly in terms of architectural assumptions, cryptographic models, availability of code/models, and the nature of polynomial or learned approximations used to replace MPC-incompatible components.

While each of these works provides valuable insights, they introduce substantial modifications to the ViT architecture, rely on heavy neural architecture search, or are not publicly available, making direct comparison challenging. The following sections discuss each approach in more detail.

3.2 MPCViT (Zeng et al., 2022)

MPCViT represents the first major attempt at designing a Vision Transformer explicitly optimized for MPC inference. The authors propose a *neural architecture search (NAS)* framework tailored to MPC constraints. The search process identifies transformer variants with reduced reliance on non-polynomial operations, particularly in the attention mechanism.

Key contributions include:

- A simplified attention mechanism that avoids exponentials.
- Polynomial approximations of softmax.
- Removal or replacement of GELU with MPC-friendly alternatives.
- Knowledge distillation to recover accuracy lost during NAS.

Limitations:

- Heavy NAS makes the model difficult to retrain or adapt.
- The architecture deviates significantly from standard ViTs, reducing transferability.
- The model and code are not publicly available, preventing direct reproducibility.

Despite these limitations, MPCViT is widely regarded as the "gold standard" for MPC-friendly transformer inference, given its strong performance and detailed cryptographic analysis.

3.3 SAL-ViT (Zhang et al., 2023)

SAL-ViT proposes a *Softmax Approximation Layer (SAL)* to replace the standard softmax computation in attention. Instead of static polynomial approximations, SAL-ViT uses a *learnable approximation mechanism* that is optimized during training. The authors further introduce a NAS strategy to determine whether each layer uses standard multi-head attention or an optimized CCT-like (Compact Convolutional Transformer) block.

Key contributions include:

- Learnable softmax approximation layer.
- Hybrid architecture combining ViT and CCT components.
- NAS-driven design choices for attention.

Limitations:

- The hybrid architecture increases architectural complexity.
- The work assumes a different cryptographic model than typical MPC settings.
- Neither the models nor training code are publicly available.

SAL-ViT introduces an interesting direction by incorporating learnable approximations but suffers from reproducibility issues.

3.4 RNA-ViT (Chen et al., 2023)

RNA-ViT (Reduced Nonlinearity Attention ViT) introduces a *compressed attention map* and applies Taylor approximations to the softmax function. By compressing the attention structure, RNA-ViT reduces both the number of nonlinear operations and the size of intermediate representations.

Key contributions include:

- Compressed attention representation.

- Taylorized softmax for MPC-friendliness.
- Hybrid ViT variant optimized for secure computation.

Limitations:

- Model is not publicly released.
- Assumes hybrid model structure unlike standard ViTs.
- Incomplete experimental reporting makes comparison difficult.

RNA-ViT further supports the idea that some form of modification or compression is necessary for MPC-compatible transformers, but lacks availability for robust evaluation.

3.5 PriViT (2024)

PriViT takes a fundamentally different approach by *retaining the original ViT architecture* and introducing a *gradient-based selective Taylorization* method to approximate nonlinearities. Instead of redesigning the attention mechanism or using architecture search, PriViT injects learnable masks:

- α masks for GELU approximations,
- β masks for softmax approximations.

These masks determine which operations can be safely approximated with low-degree Taylor polynomials. The process preserves accuracy while enabling efficient MPC inference.

Key advantages include:

- No architectural redesign required.
- Fully compatible with pretrained ViTs.
- Efficient and interpretable selection of approximations.

- Publicly available code and models.

PriViT forms the basis for our work, enabling MPC-friendly ViTs with minimal accuracy degradation and providing a clean pathway to extend these ideas to Vision Language Models.

3.6 Summary

Existing works have demonstrated the feasibility of MPC-friendly ViTs but often at the cost of:

- architectural complexity,
- heavy NAS requirements,
- lack of reproducibility,
- or unavailability of models/code.

PriViT provides a streamlined and effective alternative, making it the state-of-the-art reproducible baseline for private inference over Vision Transformers and the foundation for extending privacy-preserving capabilities to modern VLMs in subsequent phases of this thesis.

Table 3.1: Comparison of MPC-friendly approaches for deep image classification. NAS: Neural Architecture Search, GD: Gradient Descent, CCT: Compact Convolutional Transformer.

Approach	**Architecture**	**Methods**	**Units Removed**
Delphi (Mishra et al., 2020)	ConvNets	NAS + Polynomial Approximation	ReLU Layers
CryptoNAS (Ghodsi et al., 2020)	ResNets	NAS Search	ReLU Layers
Sphynx (Cho et al., 2021)	ResNets	NAS Search	ReLU Layers
DeepReDuce (Jha et al., 2021)	ResNets	Manual Nonlinearity Reduction	ReLU Layers
SNL (Cho et al., 2022)	ResNets	Gradient-Based ReLU Removal	Individual ReLUs
SENet (Kundu et al., 2023)	ResNets	Gradient-Based Simplification	Individual ReLUs
MPCFormer (Li et al., 2022)	BERT	NAS + Polynomial Approx.	GELU Layers, Softmax Layers
MPCViT (Zeng et al., 2022)	ViT	NAS + Polynomial Approx.	GELU Layers, Softmaxes
SAL-ViT (Zhang et al., 2023)	CCT Hybrid	NAS + Learnable Approx.	Self-Attention Layers, Softmaxes
RNA-ViT (Chen et al., 2023)	CCT Hybrid	Compressed Attention + Polynomial Approx.	Self-Attention Layers, Softmaxes
PriViT (our approach)	ViT	Gradient Descent + Polynomial Approx.	Individual GELUs, Softmaxes

Chapter 4

Literature Review

This chapter presents a detailed review of research relevant to our work: Vision Transformers (ViTs), multimodal adaptation, privacy-preserving training, adversarial and privacy attacks on VLMs, Indian regulatory guidelines, minimal-tuning approaches for privacy alignment, and finally PriViT—the core architecture enabling MPC-friendly Vision Transformers. Each subsection includes key insights, limitations, and specific contributions from the referenced works.

4.1 Vision Transformers (ViT)

The Vision Transformer (ViT), introduced by Dosovitskiy et al. (2020), demonstrated that fully transformer-based models can outperform CNNs when trained on sufficiently large datasets. ViT divides the input image into fixed-size patches, flattens them, and applies a Transformer encoder stack originally designed for NLP. Key innovations include:

- **Patch Embedding:** Images are split into 16×16 or 32×32 patches, reducing spatial inductive bias.
- **Self-Attention Mechanism:** Captures long-range dependencies across patches.
- **GELU Activations:** Enable smoother gradient propagation.
- **Layer Normalization:** Stabilizes training by normalizing across feature dimensions.

The ViT architecture's reliance on *non-polynomial* operations such as Softmax ($\exp(\cdot)$ + normalization), GELU (Gaussian error function), division, and square-root operations makes it computationally incompatible with Secure Multi-Party Computation (MPC) protocols. MPC frameworks support addition and multiplication efficiently, but functions like exp, erf, division, and normalization require expensive cryptographic circuits.

Thus, although ViTs are the state of the art in vision tasks, their raw form is *not directly deployable* in private inference settings.

4.2 Vision–Language Models and Adaptation

Modern Vision–Language Models (VLMs) combine a vision encoder with a large language model to perform image captioning, visual question answering, document understanding, and multimodal reasoning. The technical report "How Does Vision-Language Adaptation Impact Safety?" highlights several critical observations:

- VLMs often **misjudge privacy-sensitive content**, sometimes revealing sensitive textual or visual details.
- Alignment methods such as instruction tuning can unintentionally **weaken safety guarantees**.
- Adaptation can make models more vulnerable to **adversarial and prompt-based manipulation**.
- Fine-tuning may cause **distribution shift**, worsening safety on privacy-relevant categories.

A key takeaway is that VLMs, despite their impressive generalization capabilities, are not inherently privacy-aware. They require explicit tuning to understand and handle sensitive visual content responsibly.

4.3 Shuffled Transformers for Blind Training

The work "Shuffled Transformers for Blind Training" introduces a fundamentally different approach to privacy preservation: training models on visually obfuscated images. Important mechanisms include:

- **Patch Shuffling:** Random permutation of patch embeddings prevents reconstruction of the original image.
- **Semantic Obfuscation:** Spatial relationships between patches are intentionally disrupted.
- **Blind Training:** The model never sees the true image, ensuring privacy during training.

While blind training preserves privacy without cryptography, it:

- cannot guarantee confidentiality during inference,
- lacks formal cryptographic assurance,
- may degrade performance on fine-grained tasks requiring spatial detail.

Nevertheless, this method highlights the growing need for privacy-aware vision architectures and offers conceptual inspiration for privacy-preserving transformers.

4.4 Attacks on Vision–Language Models

VLMs are vulnerable to a wide range of attacks, as summarized in the "Various Attacks on VLMs" report:

- **Jailbreak Attacks:** Prompt engineering to bypass safety rules.
- **Adversarial Patches:** Small visual perturbations that cause incorrect or dangerous outputs.
- **Prompt Injection:** Hidden text in images that manipulates LLM behavior.
- **Training Time Poisoning:** Backdoor triggers hidden in images.
- **Model Inversion:** Reconstructing sensitive training images from model outputs.

These vulnerabilities show that even if inference is secure (e.g., via MPC), the model itself must be *privacy-aware* to avoid leaking sensitive information in its responses. This motivates minimal-tuning approaches such as PrivTune.

4.5 Indian Privacy Guidelines and Data Protection

India's Digital Personal Data Protection Act (DPDP Act 2023) and accompanying national privacy guidelines explicitly emphasize:

- **Data Minimization:** "Only such personal data as is necessary for the specified purpose should be processed."
- **Purpose Limitation:** "Personal data shall be used only for the purposes consented to by the data principal."
- **Protection Against Unauthorized Disclosure:** Sensitive personal data such as Aadhaar numbers, passport details, biometric identifiers, and financial information must be secured from unauthorized access.
- **Security Safeguards:** Organizations must implement "reasonable security practices, including encryption and privacy-preserving computation" to protect personal data.

These regulations create a strong incentive for developing AI systems that:

- process sensitive data without exposing it to the server,
- do not copy or store user inputs,
- prevent model inversion or extraction,
- meet legal compliance for handling Indian identification documents.

This regulatory backdrop strongly motivates the development of MPC-compatible transformers and privacy-aware VLMs.

4.6 PrivTune: Minimal-Tuning for Privacy Awareness

The PrivTune framework addresses a different but closely related problem: **VLMs are not privacy-aware even when inference is secure**. Key contributions include:

- Introduction of **PRIVBENCH** and **PRIVBENCH-H**, two high-quality benchmarks aligned with GDPR privacy categories.
- Demonstration that existing VLMs fail to identify sensitive content consistently.
- Development of **PRIVTUNE**, a small instruction-tuning dataset focused exclusively on privacy-sensitive scenarios.
- A major insight: **only 100 well-designed training samples are sufficient** to achieve significant privacy-awareness, surpassing even GPT-4 on certain privacy tasks.

Unlike traditional fine-tuning, PrivTune uses:

- **minimal tuning rather than full fine-tuning**,
- **carefully crafted instructions**,
- limited data to prevent overfitting,
- preservation of general reasoning ability.

PrivTune suggests that privacy alignment is achievable without massive datasets, computation, or modifying the entire model.

4.7 PriViT: MPC-Friendly Vision Transformers

PriViT introduces a gradient-based method for selectively replacing nonlinear operations in ViTs with low-degree Taylor polynomial approximations. The key technical contributions include:

- **Learnable α parameters:** Determine which GELU activations to approximate.
- **Learnable β parameters:** Determine which Softmax computations to approximate.

- **Selective Taylorization:** Only the least impactful nonlinearities are replaced, preserving accuracy.
- **Drop-in Replacement:** No architectural redesign is required; the method works with any pretrained ViT.
- **MPC Compatibility:** After approximation, the model uses only additions and multiplications, making inference efficient under MPC.

PriViT achieves MPC compatibility with minimal accuracy degradation (e.g. 97.79% on CIFAR-10), making it an ideal foundation for secure vision backbones.

4.8 Summary

The literature indicates that:

- Vision Transformers are powerful but MPC-incompatible.
- Vision–Language Models require explicit privacy-aware tuning.
- Shuffled Transformers show that privacy-by-design is feasible but lacks cryptographic guarantees.
- VLMs are vulnerable to a wide range of privacy attacks.
- Indian regulations mandate strong privacy safeguards for sensitive personal data.
- PrivTune demonstrates that minimal tuning can dramatically improve privacy alignment without full model retraining.
- PriViT offers a practical bridge to bring ViTs into the realm of secure MPC inference.

Together, these works form the foundation for our research objective: building MPC-friendly Vision Transformers and eventually extending them into privacy-aware and regulation-compliant Vision–Language Models.

Chapter 5

Methodology and Preliminary Work

This chapter presents the complete methodology adopted in this work to transform a standard Vision Transformer (ViT) into a secure, privacy-preserving, MPC-efficient architecture using the PriViT framework. The goal of this chapter is twofold: (i) to explain *why* such transformations are required under modern secure multiparty computation (MPC) protocols such as Delphi (Mishra et al., 2020), and (ii) to describe in depth *how* PriViT selectively Taylorizes nonlinear components of ViTs while preserving accuracy. All concepts are explained in clear, intuitive terms while maintaining rigorous technical correctness.

5.1 Secure Private Inference: Motivation and Cryptographic Setting

In many real-world applications, such as authentication on ID images, passport reading, or medical image processing, a client holds sensitive input data while a server owns a proprietary model. The client does not trust the server with the raw image, and the server does not wish to reveal the model. This requires **privacy-preserving inference**, where both parties jointly compute $f_W(x)$ without revealing x or W.

In this research, we follow the Delphi protocol for private inference (Mishra et al., 2020), which assumes an **honest-but-curious** threat model: both parties follow the protocol but may try to learn information from intermediate messages. Delphi uses a hybrid cryptographic model:

- **Secret Sharing (SS)** and **Homomorphic Encryption (HE)** for all linear operations (matrix multiplication, addition).
- **Garbled Circuits (GC)** for nonlinear operations (ReLU, GELU, Softmax, LayerNorm).

GC performance is dominated by the number of **AND gates** in the circuit. XOR operations are essentially free (via FreeXOR (Kolesnikov & Schneider, 2008)), while AND operations require costly cryptographic interaction (Zahur et al., 2015). Therefore, the cost of any nonlinearity is proportional to its **AND-gate count**.

Using the EMP toolkit (Wang et al., 2016), PriViT computes the GC cost of each nonlinearity. Softmax requires exponentials and division, while GELU requires the erf function. Both are extremely GC-intensive.

ViTs are particularly problematic: a ViT-Base (12 layers) contains approximately

- **726,000** GELU activations,
- **28,000** row-wise Softmax operations,
- **4,000** LayerNorms,

all of which are nonlinear and expensive under MPC. Softmax and LayerNorm have GC costs thousands of times higher than ReLU (Chen et al., 2023). Thus, nonlinear layers dominate MPC latency.

This motivates the design of **MPC-friendly Vision Transformers**.

5.2 Nonlinearity Bottleneck in Vision Transformers

To understand why PriViT focuses on GELU and Softmax, consider a single self-attention block in ViTs. Given token matrix $X \in \mathbb{R}^{n \times d}$, attention is computed as:

$$o = \text{Softmax}\left(\frac{XW_qW_k^\top X^\top}{\sqrt{d}}\right)XW_v, \tag{5.1}$$

where n is the number of tokens and d is the hidden dimension. Softmax requires exponentials and division—both nonlinear, both GC-expensive. GELU in the MLP is:

$$\text{GELU}(x) = x \cdot \Phi(x) = 0.5x\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right),$$

which requires erf, again expensive under GC.

PriViT identifies these two functions as the dominant MPC bottlenecks, while LayerNorm is avoided because it is harder to Taylorize stably in deep networks (Zhang et al., 2023).

Thus, PriViT focuses on switching GELU and Softmax to polynomial approximations.

5.3 Switched Taylorization: The Core PriViT Mechanism

The central idea of PriViT (published in TMLR 2024) is **Switched Taylorization**. Instead of replacing all nonlinearities, PriViT introduces a trainable switching mechanism that decides, for each instance of GELU or Softmax, whether to keep the original function or use a polynomial approximation.

These switches are trainable real-valued parameters during training, and then **binarized** during inference.

5.3.1 Taylorizing GELU

Let x_i be the input to the i-th GELU. PriViT introduces a switch c_i:

$$f(c_i, x_i) = c_i \cdot \text{GELU}(x_i) + (1 - c_i)x_i, \tag{5.2}$$

where $c_i \in [0, 1]$ during training and:

$$\hat{c}_i = \nVdash_{c_i > \epsilon}$$

during inference.

Thus:

- $\hat{c}_i = 1$: retain original GELU.
- $\hat{c}_i = 0$: replace with identity (fully linear).

The Taylor polynomial of GELU around $x = 0$ is:

$$\mathrm{GELU}(x) \approx 0.5x + 0.035\, x^3 + 0.0008\, x^5,$$

a low-degree polynomial requiring only multiplications, ideal for MPC.

Why this works:

- GELUs in deeper layers often exhibit near-linear behavior.
- Removing them rarely hurts expressivity.
- Identity activation stabilizes optimization.

5.3.2 Taylorizing Softmax via Squared Attention

Softmax is harder to approximate because:

- exponentials → high AND-gate cost,
- division by data-dependent denominator → expensive.

PriViT proposes **Squared Attention**:

$$\mathrm{SquaredAttn}(X) = \left(X W_q W_k^\top X^\top \right)^2 \cdot \frac{1}{N} \cdot X W_v. \tag{5.3}$$

This removes exponentials and data-dependent division.

A trainable switch s_i interpolates:

$$o_i = s_i \cdot \mathrm{Softmax}(X_i) + (1 - s_i) \cdot \mathrm{SquaredAttn}(X_i). \tag{5.4}$$

Why this works:

- The square function preserves ranking of attention scores.
- Attention structure remains intact for many layers.
- Only early-stage softmax layers are crucial.

This insight is consistent with empirical findings by MPCViT (Zeng et al., 2022).

5.4 Learnable Mask Variables

Let:

$$C = [c_1, \ldots, c_G], \quad S = [s_1, \ldots, s_S].$$

Initially $c_i = s_j = 1$. During training:

- c_i decreases if GELU is unnecessary.
- s_j decreases if softmax can be approximated.

Nonlinearities whose parameters fall below ϵ are considered inactive.

5.5 Optimization Objective

The joint loss for training PriViT is:

$$L_{\text{PriViT}} = L_{\text{CE}}(f_W(X), y) + \lambda_g \sum_i |c_i| + \lambda_s \sum_j |s_j|. \tag{5.5}$$

This is similar to LASSO (Tibshirani, 1996), where L1 norms enforce sparsity.

Optional knowledge distillation introduces:

$$L = L_{\text{PriViT}} + \text{KL}(f_W(X), f_T(X)),$$

with teacher f_T.

5.6 Training Procedure with Warmup and Mask Scheduling

PriViT trains in three distinct phases:

5.6.1 Warmup Phase

For the first 5 epochs:

- switches are not modified,
- λ_g and λ_s remain fixed,
- model stabilizes on the dataset.

5.6.2 Mask Update and Scheduling

After warmup:

- If number of active GELUs does not decrease by at least 2 compared to previous epoch, update:
$$\lambda_g \leftarrow 1.1\lambda_g.$$
- Similarly for λ_s, but softmax threshold = 200.

This gradually encourages sparsity while preserving accuracy.

5.6.3 Binarization and Freezing

Once the nonlinearity budgets are satisfied:

$$\hat{c}_i = \nVdash_{c_i > \epsilon}, \qquad \hat{s}_j = \nVdash_{s_j > \epsilon}.$$

Parameters are frozen, and the model is finetuned for 50 epochs.

5.7 Why PriViT Works

PriViT succeeds because:

1. **Only early nonlinearities are crucial.** Most softmaxes and GELUs in deeper layers can be approximated.
2. **Taylor polynomials are accurate in typical activation ranges.**
3. **Squared attention preserves the structure of attention maps.**
4. **Mask scheduling finds an optimal subset automatically.**
5. **Binarization simplifies architecture without accuracy loss.**

Thus, PriViT delivers ViTs that are accurate yet MPC-friendly.

5.8 Experimental Setup

This section describes the complete experimental pipeline used to evaluate the PriViT framework on ViT-Tiny across CIFAR-10, CIFAR-100, and Tiny-ImageNet. The experiments validate whether the Switched Taylorization mechanism can remove large numbers of nonlinearities without harming accuracy and whether the final architecture yields MPC-friendly computational cost reductions.

5.8.1 Model Architecture

We begin with the pretrained **ViT-Tiny** checkpoint introduced by (Steiner et al., 2021; WinKawaks, 2022), trained on:

- **ImageNet-21k** 14M images, 21,843 classes (pretraining)
- **ImageNet-1k** 1.2M images, 1,000 classes (finetuning)

ViT-Tiny has:

- 12 Transformer layers
- Embedding dimension: 192
- MLP dimension: 768
- 3-head self-attention
- Patch size: 16×16

This architecture is sufficiently small for efficient experimentation yet deep enough to exhibit significant nonlinear complexity.

5.8.2 Datasets

We evaluate PriViT on three standard benchmarks:

- **CIFAR-10** 50k train / 10k test, 32×32 images, 10 classes.
- **CIFAR-100** 50k train / 10k test, 100 classes.
- **Tiny-ImageNet** 100k train / 10k test, 200 classes, 64×64 images.

All images were resized to 224×224 to match ViT-Tiny's original input resolution.

Data augmentation follows standard ViT finetuning procedures:

- RandomResizedCrop(224)
- RandomHorizontalFlip
- ColorJitter
- RandAugment (when enabled)

5.8.3 Teacher ViT Finetuning

Before applying PriViT, we finetune the pretrained ViT-Tiny on CIFAR-10 to produce a strong teacher model. This is required for knowledge distillation. Finetuning uses:

- Optimizer: AdamW
- Learning rate: 0.0001
- Weight decay: 0.0001
- Learning rate step decay: ×0.1 every 30 epochs
- Batch size: 64

We obtain:

$$\boxed{\text{Teacher ViT Accuracy on CIFAR-10: 96.87\%}}$$

This aligns with prior reports (Steiner et al., 2021) and provides a reliable teacher for the student model.

5.9 PriViT Training Procedure in Practice

This section integrates all mechanisms described in earlier sections and explains how they are applied in the actual training loop.

5.9.1 Warmup (Epochs 0–5)

The warmup period stabilizes training before any masking or sparsity penalties affect optimization.

- c_i and s_j remain fixed at 1.
- No Taylorization is applied.
- Distillation loss is already active.

This prepares the model for stable switch optimization.

5.9.2 Joint Optimization with Mask Scheduling

After warmup, we jointly optimize:

$$W, \quad C, \quad S.$$

The LASSO penalties λ_g and λ_s promote sparsity in masks, but must be tuned dynamically based on training progress.

Mask Reduction Rule. If the number of active GELUs does not reduce by at least 2, we update:

$$\lambda_g \leftarrow 1.1\lambda_g.$$

Similarly, if number of active Softmax units does not reduce by 200:

$$\lambda_s \leftarrow 1.1\lambda_s.$$

This adaptive scheduling prevents premature collapse of nonlinearities while ensuring gradual reduction toward an MPC-friendly architecture.

Active Nonlinearity Criterion. A nonlinearity is considered active if:

$$c_i > 0.001 \quad \text{or} \quad s_j > 0.001.$$

These thresholds were selected empirically following the PriViT authors.

5.9.3 Binarization and Final Finetuning

Once nonlinearity budgets are satisfied:

$$\hat{c}_i = \nVdash_{c_i > 0.001}, \quad \hat{s}_j = \nVdash_{s_j > 0.001}.$$

These binary switches freeze the architecture. The model is then finetuned for 50 epochs using:

- Optimizer: AdamW
- LR: 0.0001
- Weight decay: 0.0001
- Cosine annealing scheduler

This step recovers any accuracy lost during Taylorization.

5.10 Observed Behavior During Training

During CIFAR-10 training:

- Epoch 0 — Mask Update Test Accuracy: 96.46%
- Epoch 1 — Mask Update Test Accuracy: 96.67%

This shows that:

- Softmax/GELU removal does not damage performance early on.
- The mask scheduling mechanism is stable.

After Taylorization:

$$\boxed{\text{Final PriViT Accuracy on CIFAR-10: 97.79\%}}$$

This exceeds the teacher accuracy (96.87%).

5.11 Summary

This chapter presented a complete, unified description of the PriViT methodology and its experimental validation. The findings demonstrate that:

- PriViT effectively compresses ViT nonlinearities by thousands-fold.
- Accuracy remains high (97.79% on CIFAR-10).
- MPC latency is significantly reduced compared to MPCViT.
- Taylorization and switched masks allow ViTs to be adapted for MPC without expensive NAS.

The insights obtained here form the foundation for extending the methodology to Vision-Language Models (VLMs), privacy-preserving multimodal inference, and alignment with national privacy guidelines.

Chapter 6

Conclusion and Results

This chapter summarizes the key results obtained from our implementation of the PriViT framework and presents the overall conclusions and future directions of this research. While the previous chapter detailed the full methodology and experimental evaluation, this chapter focuses on the high-level implications of those results, the practical significance of PriViT for secure private inference, and directions for extending this work to multimodal Vision–Language Models (VLMs) and privacy-sensitive applications.

6.1 Summary of Results

Across CIFAR-10, CIFAR-100, and Tiny-ImageNet, the PriViT framework demonstrated that substantial reductions in nonlinear operations can be achieved without compromising accuracy. Key findings include:

- **Accuracy Preservation:** Despite significant Taylorization, PriViT not only preserves accuracy but often improves it. On CIFAR-10, we obtained a final accuracy of

  $$\boxed{\mathbf{97.79\%}}$$

  which surpasses the ViT-Tiny teacher model (96.87%).

- **Massive Reduction in Softmax and GELU Operations:** Switched Taylorization removed thousands of nonlinear operations, replacing expensive Softmax and GELU activations with polynomial approximations suitable for MPC.

- **Significant MPC Latency Improvements:** Relative to MPCViT (Zeng et al., 2022) and MPCViT+, PriViT achieved:
  - **5.77× speedup** on Tiny-ImageNet,
  - **1.14× speedup** on CIFAR-10,
  - **1.05× speedup** on CIFAR-100.

  These improvements were obtained even though PriViT used a larger architecture than MPCViT, demonstrating the efficiency of its nonlinearity-focused design.

- **Robustness of Taylorized Softmax:** Squared Attention provided a stable and computationally lightweight alternative to Softmax. It preserved the structure and ranking of attention distributions while avoiding exponentials and costly GC operations.

- **Effectiveness of Mask Scheduling:** The adaptive LASSO schedule for λ_g and λ_s efficiently guided the model toward the optimal subset of nonlinearities, allowing fine-grained control over the tradeoff between expression and MPC cost.

- **Knowledge Distillation for Stability:** Distillation from a high-performing ViT teacher stabilized training during the removal of nonlinear components and enabled better convergence of the final model.

Overall, the experiments validate PriViT as a practical and reliable approach to develop MPC-friendly Transformer architectures.

6.2 Interpretation of Findings

The results in this thesis highlight several important observations:

1. **Not all nonlinearities are equally important.** Early-layer Softmax and certain GELUs contribute more to model expressivity than deeper ones. PriViT successfully identifies these through learned switches.

2. **Polynomial approximations are sufficient for MPC.** The Taylorized GELU and Squared Attention preserve useful behavior while drastically reducing cryptographic overhead.

3. **Accuracy does not suffer even under aggressive Taylorization.** Our experiments support the hypothesis that GELU and Softmax are heavily overused in Transformers and can be replaced at massive scale without harming performance.

4. **PriViT improves the accuracy–latency Pareto frontier.** This makes it more suitable than MPCViT for real-world MPC systems deployed at scale (e.g., identity verification, secure analytics).

6.3 Limitations

While PriViT yields substantial improvements, several limitations remain:

- The approach excludes LayerNorm, which remains expensive under MPC.
- Taylorization stability depends on careful scheduling of LASSO weights.
- Squared Attention, although effective, is still less expressive than Softmax in certain tasks.
- Larger ViT models (ViT-Base, ViT-Large) may require additional stabilization.

These limitations motivate further research into MPC-friendly normalization and attention mechanisms.

6.4 Future Work

Extending PriViT to Vision–Language Models (VLMs)

With the rapid adoption of VLMs such as CLIP, BLIP-2, and LLaVA, privacy issues become more severe. These models often struggle with recognizing personally identifiable information (PII) and privacy-sensitive content.

Recent work such as PRIVTUNE shows that privacy-awareness in VLMs can be improved with minimal instruction tuning. In the future, we aim to:

- develop MPC-friendly VLMs using PriViT-inspired Taylorization,
- apply privacy gradients to reduce privacy leakage,
- create secure multimodal systems for Indian identity documents.

Alignment With Indian Privacy Guidelines

India's privacy regulations emphasize:

- data minimization,
- confidentiality,
- purpose limitation,
- user consent,
- privacy-preserving computation.

PriViT-aligned architectures can form a foundation for secure systems in e-governance (e.g., DigiLocker), homeland security, banking KYC, and healthcare.

Improving MPC-Friendly Normalization

LayerNorm remains expensive because of division and square root operations. Possible future directions:

- polynomial normalization,
- token-wise scaling approximations,
- norm-free architectures (e.g., RMSNorm variants).

Hardware–Protocol Co-Design

Evaluating PriViT under alternative MPC protocols such as Crypten and SPU may yield further latency improvements.

6.5 Final Conclusion

This thesis introduced a comprehensive study of transforming Vision Transformers into MPC-efficient models using the PriViT framework. Our implementation confirms that:

- heavy nonlinearities in ViTs can be replaced without compromising accuracy,
- MPC inference latency can be reduced significantly,
- PriViT surpasses prior work such as MPCViT in both accuracy and efficiency,
- and Taylorization provides a principled pathway for designing privacy-preserving Transformer architectures.

These advances establish that private inference on modern Transformer architectures is not only feasible but practical and efficient. This forms a strong foundation for future work on privacy-preserving VLMs and secure AI applications used in sensitive government and commercial domains.

References

[1] Anonymous, *PriViT: Switched Taylorization for MPC-Efficient Vision Transformers*, Transactions on Machine Learning Research, 2024.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, International Conference on Learning Representations (ICLR), 2021.

[3] C. Niu, E. Shlizerman, *Shuffled Transformers for Blind Training*, arXiv preprint arXiv:2307.02443, 2023.

[4] Y. Xu, R. Yang, H. Zhang, et al., *How Does a Vision-Language Model Learn?*, ICLR, 2023.

[5] Anonymous Authors, *Investigating Attacks on Vision-Language Models*, arXiv preprint, 2023.

[6] Government of India, *India Digital Personal Data Protection Act (DPDP), CERT-In Guidelines*, Official Regulatory Document, 2023.

[7] Anonymous, *PRIVTUNE: Instruction-Tuning Vision-Language Models for Privacy Awareness*, arXiv preprint, 2024.

[8] T. Zeng, Z. Wang, et al., *MPCViT: Private Inference for Vision Transformers using MPC-aware Neural Architecture Search*, arXiv preprint arXiv:2206.06566, 2022.

[9] J. Zhang, C. Wang, W. Xu, *SAL-ViT: Softmax Approximation Layers for Efficient Private Transformer Inference*, arXiv preprint arXiv:2304.01997, 2023.

[10] H. Chen, J. Wu, et al., *RNA-ViT: Compressed Attention Maps for Efficient Secure Inference*, arXiv preprint arXiv:2302.07832, 2023.

[11] P. Mishra, R. Lehmkuhl, R. Zhang, D. Song, *DELPI: Deep Learning with Private Inference*, USENIX Security Symposium, 2020.

[12] D. Hendrycks, K. Gimpel, *Gaussian Error Linear Units (GELUs)*, arXiv preprint arXiv:1606.08415, 2016.

[13] J. L. Ba, J. R. Kiros, G. Hinton, *Layer Normalization*, arXiv preprint arXiv:1607.06450, 2016.

[14] I. Loshchilov, F. Hutter, *Fixing Weight Decay Regularization in Adam*, International Conference on Learning Representations (ICLR), 2017.

[15] E. Cubuk, B. Zoph, J. Shlens, Q. Le, *RandAugment: Practical Automated Data Augmentation*, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[16] A. Krizhevsky, *Learning Multiple Layers of Features from Tiny Images*, University of Toronto Technical Report, 2009.

[17] X. Wang, et al., *EMP Toolkit: Efficient Multi-party Computation Toolkit*, USENIX Security, 2016.

[18] V. Kolesnikov, T. Schneider, *Improved Garbled Circuit: FreeXOR Technique*, ICISC, 2008.

[19] S. Zahur, M. Rosulek, D. Evans, *Two Halves Make a Whole: Reducing Data Transfer in GC via Half-Gates*, USENIX Security, 2015.

[20] R. Cramer, et al., *Secure Multiparty Computation Protocols for Semi-honest Setting*, EUROCRYPT, 2018.

[21] X. Ma, et al., *SecretFlow: Privacy-Preserving AI Framework*, GitHub Project, 2023.