RESEARCH-ARTICLE

Rhombus: Fast Homomorphic Matrix-Vector Multiplication for Secure Two-Party Inference

JIAXING HE, Ant group, Hangzhou, Zhejiang, China

KANG YANG

GUOFENG TANG, Ant group, Hangzhou, Zhejiang, China

ZHANGJIE HUANG, Ant group, Hangzhou, Zhejiang, China

LI LIN, Ant group, Hangzhou, Zhejiang, China

CHANGZHENG WEI, Ant group, Hangzhou, Zhejiang, China

View all

Open Access Support provided by:

Ant group

Rhombus: Fast Homomorphic Matrix-Vector Multiplication for Secure Two-Party Inference

Jiaxing He
Digital Technologies, Ant Group
Beijing, China
jiaxing.hjx@antgroup.com

Kang Yang*
State Key Laboratory of Cryptology
Beijing, China
yangk@sklc.org

Guofeng Tang
Digital Technologies, Ant Group
Beijing, China
tang.guofeng789@gmail.com

Zhangjie Huang
Digital Technologies, Ant Group
Hangzhou, China
zhangjie.hzj@antgroup.com

Li Lin
Digital Technologies, Ant Group
Beijing, China
felix.ll@antgroup.com

Changzheng Wei
Digital Technologies, Ant Group
Shanghai, China
changzheng.wcz@antgroup.com

Ying Yan
Digital Technologies, Ant Group
Beijing, China
fuying.yy@antgroup.com

Wei Wang
Digital Technologies, Ant Group
Hangzhou, China
wei.wangwwei@antgroup.com

ABSTRACT

We present *Rhombus*, a new secure matrix-vector multiplication (MVM) protocol in the semi-honest two-party setting, which is able to be seamlessly integrated into existing privacy-preserving machine learning (PPML) frameworks and serve as the basis of secure computation in linear layers. *Rhombus* adopts RLWE-based homomorphic encryption (HE) with coefficient encoding, which allows messages to be chosen from not only a field $\mathbb{F}_p$ but also a ring $\mathbb{Z}_{2^\ell}$, where the latter supports faster computation in non-linear layers. To achieve better efficiency, we develop an input-output packing technique that reduces the communication cost incurred by HE with coefficient encoding by about 21×, and propose a split-point picking technique that reduces the number of rotations to that sublinear in the matrix dimension. Compared to the recent protocol *HELiKs* by Balla and Koushanfar (CCS'23), our implementation demonstrates that *Rhombus* improves the whole performance of an MVM protocol by a factor of 7.4× ∼ 8×, and improves the end-to-end performance of secure two-party inference of ResNet50 by a factor of 4.6× ∼ 18×.

CCS CONCEPTS

• **Theory of computation → Cryptographic protocols**.

KEYWORDS

Two-party computation; Secure inference; Homomorphic matrix multiplication; Coefficient encoding

*Corresponding author

1 INTRODUCTION

Machine learning (ML) has enabled numerous applications. Particularly, model owners can provide inference services for users without the capability of training models, i.e., so-called ML as a Service (MLaaS). Users can obtain value from ML services, while model owners are able to effectively monetize their services. However, MLaaS puts forward a challenging question regarding data privacy, i.e., a model owner may learn private data of a user, or the user may reveal the model. Privacy-preserving machine learning (PPML) addresses the challenging question by providing a solution of secure two-party inference (i.e., two parties can collaboratively perform secure inference without revealing data privacy).

Homomorphic encryption (HE), a powerful cryptographic primitive, enables the model owner to perform complex computations over the encrypted user data. Although many efforts [34, 36, 51] have been made to execute the HE-based computations efficiently, it's still impractical to perform all the inference computations on encrypted data because the model owner has to invoke bootstrapping operations, which are extremely computation-intensive. Instead, many protocols (e.g., [1, 5, 9, 10, 14, 18, 21, 23, 27, 28, 32, 35, 37, 40–42, 44–48, 50, 52, 54]) have been developed to improve the efficiency of PPML under the hybrid framework combining the HE and secure multi-party computation (MPC) techniques. An ML model consists of a sequence of linear and non-linear layers. Specifically, the operations in linear layers, e.g., 3D convolution and fully connection (FC), can be often modeled as matrix-vector multiplications (MVMs), which can be realized from HE based on the ring learning with errors (RLWE) [39], as that of in prior works [10, 27, 32, 40, 41, 48], while the computations in non-linear layers are performed by MPC

based on secret sharings. As a result, in this hybrid HE-MPC framework, each time when performing the MVM using HE, the encrypted results need to be converted to secret sharings for the subsequent computations in non-linear layers, and thus, the multiplicative depth is only one and bootstrapping is not needed. Figure 1 presents a HE-based two-party MVM paradigm, where the model owner (Bob) holds a matrix $\mathbf{W}$ while the user (Alice) holds a vector $\mathbf{v}$, and the protocol outputs a vector of additive secret sharings $(\mathbf{r}_1, \mathbf{r}_0)$ on $\mathbf{W} \cdot \mathbf{v}$ such that $\mathbf{r}_0 + \mathbf{r}_1 = \mathbf{W} \cdot \mathbf{v}$.

Alice(v)	**Bob(W)**
Encrypt $\mathbf{v}$	
$\xrightarrow{\text{Enc}(\mathbf{v})}$	
	$\text{Enc}(\mathbf{Wv}) \leftarrow \mathbf{W} \cdot \text{Enc}(\mathbf{v})$
	Sample $\mathbf{r}_1$ at random
	$\text{Enc}(\mathbf{r}_0) \leftarrow \text{Enc}(\mathbf{Wv}) - \mathbf{r}_1$
$\xleftarrow{\text{Enc}(\mathbf{r}_0)}$	
Decrypt $\mathbf{r}_0$	Decode $\mathbf{r}_1$
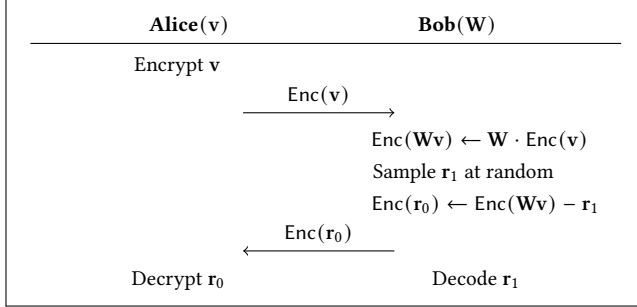
Figure 1: Secure two-party MVM paradigm based on HE.

There are two kinds of encoding approaches for RLWE-based HE to realize the MVM protocol: (1) NTT encoding that applies number theoretic transform (NTT) to the original data and then maps them to the coefficients of the plaintext polynomial, and (2) coefficient encoding that directly maps the original data to the coefficients of the plaintext polynomial. Most of prior works [5, 32, 43, 48, 58] designed HE-based MVM protocols with NTT encoding. However, NTT encoding has three drawbacks. Firstly, the NTT encoding is more costly compared to coefficient encoding. Secondly, NTT encoding is limited to operating over a field $\mathbb{F}_p$. As shown in [48], the communication cost of the oblivious transfer (OT) based protocols in non-linear layers over $\mathbb{F}_p$ is approximately 1.5× larger than that over a ring $\mathbb{Z}_{2^\ell}$ where $p \approx 2^\ell$. Thirdly, when converting homomorphic-computation results to additive secret sharings, larger HE parameters must be chosen to protect the privacy of secret values used in the homomorphic operations. The recent PPML protocol named *Cheetah* [27] employs coefficient encoding to overcome these drawbacks, at the cost of introducing a one-bit error to the computational result, which has no impact on the inference accuracy as the result would always be truncated. However, *Cheetah* encodes the MVM or 3D-convolution result in multiple ciphertexts (each encodes only a few elements of the results), which brings about a large communication overhead. In particular, the pointwise convolutions (with 1×1 filter size) incur a large communication overhead for *Cheetah*, accounting for 80% overhead of all convolutions for secure inference of ResNet50 [22], a popular deep neural network (DNN). This leads to a larger communication in linear layers of *Cheetah*, compared to the state-of-the-art PPML protocol [5] using HE with NTT encoding.

As analyzed above, we focus on coefficient encoding in this work. In the cleartexts, the MVM could be viewed as the inner products between each row of the matrix and the vector. The computation of each inner product is realized via placing the elements of matrix's

row and vector into two separate polynomials, and then multiplying them. Based on this, one naive approach comes into play: Alice encodes her vector $\mathbf{v}$ into the coefficients of a polynomial in sequence, encrypts it, and then sends the ciphertext to Bob. After receiving the ciphertext, Bob encodes each row of the matrix into the coefficients of a plaintext polynomial in some reversed order, and multiplies the encrypted vector by each encoded row. As a result, Bob generates r encrypted polynomials for a $r \times c$ matrix. To achieve $O(1)$ communication complexity, Bob packs the r ciphertexts into a single one by invoking the packing technique, known as PackLWEs [11], at the cost of $r - 1$ homomorphic rotations. In conclusion, the naive approach requires $O(r)$ plaintext-ciphertext multiplications and rotations.

1.1 Our Contribution

The naive approach, as described above, is inefficient when handling the matrices with a moderately large number r of rows (that is the case for most popular ML models), even if the number of columns c is small. In this paper, we propose two techniques to significantly reduce the computation cost, while keeping the communication complexity of $O(1)$. On one hand, we develop an *input-output packing* technique for MVM to reduce the number of rotations and plaintext-ciphertext multiplications to $O(rc/N)$, where r, c ($\leq N$) are the numbers of rows and columns in a matrix, and N is the dimension of a polynomial ring used in RLWE-based HE. On the other hand, we present a *split-point picking* technique, which further reduces the number of rotations to $O(\sqrt{rc/N})$. This results in a lower computational cost for an end-to-end execution of the MVM protocol, compared to previous protocols.

Building upon the above techniques and RLWE-based HE with coefficient encoding, we design *Rhombus*, a new two-party matrix-vector multiplication (MVM) protocol with semi-honest security. When applying it to secure matrix multiplication, we further present two new encoding methods to reduce the communication overhead, at the cost of sightly larger computational cost. The two encoding methods enable *Rhombus* to better accommodate matrices of various dimensions. Our protocol *Rhombus* accepts the HE plaintexts from not only a field $\mathbb{F}_p$ but also a ring $\mathbb{Z}_{2^\ell}$. As a building block, *Rhombus* is able to be seamlessly integrated into existing two-party PPML frameworks to improve the performance of linear layers from simple models (e.g., logistic regression) to complicated models (e.g., Transformer-based models like GPT [8]). For performance comparison between *Rhombus* and the state-of-the-art protocols [5, 27], we evaluate on the popular DNN model ResNet50 and use it as an example, as in the recent work [5].

We implemented *Rhombus* and integrated it to the open-sourced PPML library [26]. Compared to the recent protocol *HELiKs* [5] (CCS 2023) using HE with NTT encoding, *Rhombus* improves the whole performance (resp., communication cost) of an MVM protocol by 7.4× ∼ 8× (resp., 1.8×), and improves the end-to-end performance of secure ResNet50 inference by 4.6× ∼ 18×. Compared to the recent protocol *Cheetah* [27] (USENIX Security 2022) using HE with coefficient encoding, *Rhombus* reduces the communication cost of securely computing a matrix-vector multiplication (resp., pointwise convolutions) by 21× (resp., 4.6×).
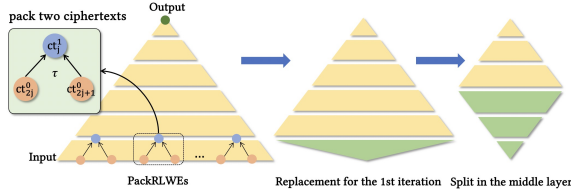
Figure 2: The high level idea of split-point picking.

1.2 Technical Overview

We adopt the HE-based two-party MVM paradigm described in Figure 1, where Alice (resp., Bob) holds a vector $\mathbf{v}$ (resp., matrix $\mathbf{W}$) over an arithmetic domain (e.g., $\mathbb{F}_p$ or $\mathbb{Z}_{2^\ell}$), and $\mathbf{W} \cdot \mathbf{v}$ is converted to a vector of additive secret sharings $(\mathbf{r}_0, \mathbf{r}_1)$ such that $\mathbf{r}_0 + \mathbf{r}_1 = \mathbf{W} \cdot \mathbf{v}$. All secret values are encrypted with HE, and the computation is performed homomorphically. In this work, we first introduce two techniques, i.e., *input-output packing* and *split-point picking*, to improve homomorphic matrix-vector multiplication (MVM), which reduce the number of rotations to sublinear level to the dimension of the matrix, while keeping the communication complexity at $O(1)$. Then, we extend MVM to securely compute multiplication of two matrices by describing two new encoding methods to reduce the communication.

Matrix vector multiplication. Matrix vector multiplication can be regarded as either the inner product of each row of a matrix with a vector (called row-major), or as the linear combination of the matrix's columns with respect to the vector (called column-major). From these two perspectives, we respectively utilize the PackLWEs [11] and Expand [2, 3] algorithms to design the HE-based MVM protocol, where the algorithms will be explained below. Suppose that a matrix $\mathbf{W}$ is of dimension $r \times c$, and r, c are powers of 2 for simplicity. In the row-major approach, Bob has to compute the inner product between each row of $\mathbf{W}$ and $\mathbf{v}$ homomorphically, which can be realized from the implicit convolution of the coefficients when multiplying two polynomials, as in [27]. Specifically, Bob encodes each row of $\mathbf{W}$ as a polynomial $\widehat{w}_i$ for each $i \in \{0, 1, \dots, r-1\}$, and performs plaintext-ciphertext multiplication to obtain r ciphertexts, where the constant terms of the corresponding plaintext polynomials are the results of inner product. Moreover, Bob can pack N/c rows into a length-N vector when $c < N$ (where N is the ring dimension in HE), and then computes N/c inner products from a single plaintext-ciphertext multiplication. To compress these ciphertexts, Bob first extracts the corresponding r LWE ciphertexts that encrypt r inner products from $m = rc/N$ RLWE ciphertexts ct_i for $i \in \{0, 1, \dots m - 1\}$, and then runs PackLWEs to merge the r LWE ciphertexts into a single RLWE ciphertext, as in [49].

We make two key improvements to the approach described as above. We first observe that the LWE extraction is unnecessary because the results of inner products are located at indices of multiples of c within each coefficient vector of the plaintext polynomials corresponding to ct_i for $i \in \{0, 1, \dots, m - 1\}$, where c is a power of two. We can modify the PackLWEs function to directly *extract-and-merge* the coefficients at powers-of-two positions of each plaintext polynomial into a single RLWE ciphertext, where we call the procedure as PackRLWEs (see Section 3.1 for details). In this way, only

$O(rc/N)$ of homomorphic rotations are needed, compared to $O(r)$ in [49]. Then, a more important observation is that the input ciphertexts to PackRLWEs are obtained by multiplying different plaintext polynomials $\{\widehat{y}_i\}$ (each packing N/c rows of matrix $\mathbf{W}$) with the same ciphertext $[\![\mathbf{v}]\!] = \mathsf{Enc}(\mathbf{v})$. In other words, the input ciphertexts of PackRLWEs are all in the form of $\widehat{y}_i \cdot [\![\mathbf{v}]\!]$, and we can try to separate the plaintexts from the ciphertexts. Then the homomorphic rotation (a.k.a., automorphism) in PackRLWEs can be performed on the common ciphertext $[\![\mathbf{v}]\!]$, and the rotation results can be reused.

To be specific, we delve into more details about PackRLWEs. Its computation pipeline is structured as a full binary tree (depicted in the left triangle of Figure 2), where each leaf node represents an input ciphertext, the internal nodes represent the intermediate ciphertexts during the computation, and the root node represents the output ciphertext. The ciphertext of each internal node is obtained via running the *pack two ciphertexts* procedure on the inputs of its two child ciphertexts and an automorphism τ, illustrated in the upper left corner of Figure 2 (see Algorithm 1 of Section 3.1 for details). Note that the ciphertexts at the same level (with the exception of those on the leaf nodes) of the binary tree are computed using the same automorphism τ. Given m ciphertexts as input, PackRLWEs processes through $\log m$ iterations, with the number of ciphertexts halving in each iteration, until only one ciphertext left (as the output). Below, we set $ct_i^0 := ct_i$, where the superscript indicates the index of the iteration in which the ciphertext resides. Specifically, in the 1st iteration, the m ciphertexts ct_i's are fed into PackRLWEs, where every two adjacent ciphertexts will be packed to one ciphertext using the *pack two ciphertexts* procedure. At the 1st layer (i.e., the 1st iteration), this procedure executes as follows:

$$ct_j^1 = \left(ct_{2j}^0 + X^{c/2} \cdot ct_{2j+1}^0\right) + \tau\left(ct_{2j}^0 - X^{c/2} \cdot ct_{2j+1}^0\right)$$

for $j \in \{0, 1, \dots, m/2 - 1\}$, and requires $N/2$ homomorphic automorphism calls, where τ satisfies $\tau(X^{c/2}) = -X^{c/2}$. By replacing ct_i^0 with $\widehat{y}_i \cdot [\![\mathbf{v}]\!]$, we derive the following equation:

$$ct_j^1 = \left(\widehat{y}_{2j} + X^{c/2} \cdot \widehat{y}_{2j+1}\right) \cdot [\![\mathbf{v}]\!] + \left(\tau(\widehat{y}_{2j}) + X^{c/2} \cdot \tau(\widehat{y}_{2j+1})\right) \cdot \tau([\![\mathbf{v}]\!]).$$

This equation implies that only one homomorphic automorphism is needed now, since $\tau([\![\mathbf{v}]\!])$ could be reused for the computation of all ct_j^1's. At this point, the computational flow has transformed into the shape like the middle part of Figure 2. Subsequently, we can move on to perform the replacement recursively to the 2nd, 3rd,... iterations, and will achieve the minimum number of homomorphic automorphism calls when we arrive the middle layer of the PackRLWEs tree (see Section 3.2 for details). At this time, we obtain a *rhombus-shaped computation flow* for homomorphic automorphism, as shown in the right part of Figure 2. In particular, the number of homomorphic automorphism calls becomes $1 + 2 + \dots + 2^{\lfloor \log m/2 \rfloor - 1} = 2^{\lfloor \log m/2 \rfloor} - 1$ for the first $\lfloor \log m/2 \rfloor$ iterations, and $(2^{\lceil \log m/2 \rceil - 1} + \dots + 2 + 1) = 2^{\lceil \log m/2 \rceil} - 1$ for the last $\lceil \log m/2 \rceil$ iterations. Totally, $2^{\lceil \log m/2 \rceil} + 2^{\lfloor \log m/2 \rfloor} - 2$, that is, $O(\sqrt{m}) = O(\sqrt{rc/N})$ of homomorphic automorphism calls are required now. During the process, we pick the optimal split-point to switch the algorithm between the replacement and PackRLWEs. We refer to the above technique as *split-point picking* (SPP).

Our column-major based approach leverages the Expand algorithm, following the combination of columns of the matrix with the

vector. Specifically, $\mathsf{Expand}(\llbracket\mathbf{v}\rrbracket)$ can output c ciphertexts $\llbracket\mathbf{v}[i]\rrbracket$ on the i-th element of $\mathbf{v}$ for $i \in \{0, 1, \ldots, c-1\}$. Hence, we encode each column of $\mathbf{W}$, multiply them with $\llbracket\mathbf{v}[i]\rrbracket$, and then accumulate the computational results. This process can also benefit from the above SPP approach, where we can similarly perform the replacement as in the row-major approach, and reduce the number of homomorphic automorphisms calls to $O(\sqrt{c})$. The column-major approach is less efficient than the row-major approach, as the latter can benefit from the input-packing. However, the column-major approach will be more efficient for tall matrices with large dimensions (i.e., $r > c \geq N$), due to the choice of split point during the SPP procedure. We postpone the explanation to Section 3.3.

Matrix Multiplication. Assume that Alice has a matrix $\mathbf{Y}_{m \times k}$ (to be encrypted), while Bob holds another matrix $\mathbf{X}_{n \times m}$. Many prior works such as [5, 27, 32] realized the matrix multiplication from matrix vector multiplication directly. However, the direct approach incurs significant communication overhead when m is small. To solve the problem, we give two new matrix-encoding methods called V1 and V2, and both of them are more communication efficient than prior approaches. The intuition in V1 (resp., V2) is to pack as many columns (resp., rows) of matrix $\mathbf{Y}_{m \times k}$ as possible into one ciphertext. Then, we can follow the row-major approach of MVM as described above, with the only difference where we have essentially swapped the roles of the matrix and vector, i.e., the matrix is now encrypted, while the vector (each row of $\mathbf{X}_{n \times m}$) is in plaintext to perform homomorphic computation. We refer the reader to Section 4.1 for details.

1.3　Related Work

The HE-based MVM protocol could be traced back to Halevi and Shoup's work [19] (*HS* in short), which has been implemented in the open-sourced library *HELib* [20]. They encode a matrix diagonally such that the output is aligned and the encrypted vector is rotated by all steps, and then multiply each encoded diagonal with the corresponding rotated vector and accumulate the multiplication results. *GAZELLE* [32] first applied the method to secure two-party inference of DNN. Specifically, they proposed a hybrid method with linear complexity to the matrix dimension based on the diagonal encoding. Based on the approach of *GAZELLE*, *GALA* [58] deferred the final *rotate-and-sum* operation to be performed on cleartext. Very recently, *HELiKs* [5] further presented a MAC (multiply-accumulate) mode, such that the rotation operations are performed after plaintext-ciphertext multiplications, which reduces the noise growth. All the above HE-based protocols adopt NTT encoding.

For coefficient encoding, *Cheetah* [27] computes the inner product from the implicit negacyclic convolution computation on the coefficients when multiplying two polynomials. However, their encoding method results in the output of multiple ciphertexts. To compress the results, Ren et al. [49] further extract specific positions of the RLWE ciphertexts to multiple LWE ciphertexts and then call the PackLWEs [11] algorithm to merge these LWE ciphertexts back into a single RLWE ciphertext. Unfortunately, the invocation of PackLWEs incurs significant computation overhead, and they accelerated it with FPGA hardware to achieve a performance that is comparable with *Cheetah*. We refer the reader to Section 5.1 for

the efficiency comparison between *Rhombus* and previous MVM protocols using HE with NTT/coefficient encoding.

Matrix multiplication can be naturally generalized from matrix-vector multiplication, which is done in *GAZELLE*, *Cheetah* and *HELiKs*. Nevertheless, the generalization may incur a significant amount of redundant computation and communication for some dimensions of matrices. Specifically, for two matrices of respective dimension $n \times m$ and $m \times k$, it directly increases the communication and computation overhead by a factor of k with respect to the matrix-vector multiplication with parameters (n, m). Several recent works devote to solving this issue to a certain extent in the context of secure two-party inference of Transformers. In particular, *Iron* [21] generalizes the approach in *Cheetah*, adopts the coefficient encoding without being confined to column-wise partitioning of the second matrix, and achieves communication complexity of $O(\sqrt{nmk/N})$. Based on *Iron*, *BumbleBee* [38] further compresses the output ciphertexts with the proposed *interleaving* packing technique, but requires a large number of rotations. Very recently, *BOLT* [43] generalizes the *GAZELLE*'s encoding method, makes full use of each slot of the polynomial, and achieves the optimal communication complexity (i.e., $O(k(m+n)/N)$). However, it needs to run an additional protocol to convert the computation results from a field $\mathbb{F}_p$ to a ring $\mathbb{Z}_{2^\ell}$ for every execution of secure matrix multiplication, due to the use of NTT encoding. See Section 5.2 for efficiency comparison between *Rhombus* and prior matrix-multiplication protocols.

In addition, there are a few works [25, 31, 57] that focus on matrix multiplication in the context of outsourced computation. In such context, both matrices need to be encrypted, and the resulting ciphertexts are involved in subsequent homomorphic computations (rather than being converted into secret sharings). Therefore, the encoding method must be compatible with subsequent computations. This brings about a large number of multiplications and rotations, resulting in a low efficiency when being applied in the two-party computation scenarios.

2　PRELIMINARIES

2.1　Notation

We use $[n]$ to denote the set $\{0, 1, \cdots, n-1\}$. A vector (resp., matrix) is represented by a bold lower-case (resp., upper-case) letter such as $\mathbf{a}$ (resp., $\mathbf{W}$). The j-th element of vector $\mathbf{a}$ is denoted by $\mathbf{a}[j]$, and the i-th row vector and j-th column vector of a matrix $\mathbf{W}$ are denoted by $\mathbf{W}_{i:}$ and $\mathbf{W}_{\cdot j}$, respectively. For two vectors $\mathbf{a}, \mathbf{b}$, we use $\langle \mathbf{a}, \mathbf{b} \rangle$ to denote the inner product of $\mathbf{a}$ and $\mathbf{b}$. We use lower-case letters with a "hat" symbol such as $\hat{a}$ to represent a polynomial, and $\hat{a}[j]$ to denote the j-th coefficient of $\hat{a}$. For a polynomial $\hat{a}$ (resp., a vector $\mathbf{a}$), we use $\llbracket\hat{a}\rrbracket$ (resp., $\llbracket\mathbf{a}\rrbracket$) to denote its encryption. For a set (resp., distribution) $\mathcal{D}$, $a \leftarrow_\$ \mathcal{D}$ represents that a is sampled uniformly from $\mathcal{D}$ (resp., according to the distribution $\mathcal{D}$).

2.2　Cyclotomic Field

Let $\zeta = \exp(\pi i / N)$ for a power-of-two integer N, then $K = \mathbb{Q}[\zeta]$ is the $2N$-th cyclotomic field and $R = \mathbb{Z}[\zeta]$ is the ring of integers of K. We will identify K (resp., R) with $\mathbb{Q}[X]/(X^N+1)$ (resp., $\mathbb{Z}[X]/(X^N+1)$) with respect to the map $\zeta \mapsto X$. The residue ring of R modulo an integer q is denoted by $R_q = R/qR$. An element of K (resp., R, R_q)

can be uniquely represented as a polynomial of degree less than N with coefficients in $\mathbb{Q}$ (resp., $\mathbb{Z}, \mathbb{Z}_q$), such as $\widehat{a}(X) = \widehat{a}[0] + \widehat{a}[1] \cdot X + \cdots + \widehat{a}[N-1] \cdot X^{N-1}$. To extend, we have the following field extension chain:

$$\mathbb{Q} = K_0 \le K_1 \le K_2 \le \cdots \le K_{\log N} = K$$

where K_ℓ denotes the $2^{\ell+1}$-th cyclotomic field. In this work, each element $\widehat{a} \in K_\ell$ is represented as a polynomial with only non-zero terms indexed by multiples of $N/2^\ell$, with $\widehat{a} = \sum_{i=0}^{2^\ell - 1} \widehat{a}[i] \cdot X^{iN/2^\ell}$.

2.3 Galois Group and Automorphisms

We recall that $K \ge \mathbb{Q}$ is a Galois extension and its Galois group $\text{Gal}(K/\mathbb{Q})$ consists of the automorphisms $\tau_d : \widehat{a}(X) \mapsto \widehat{a}(X^d)$ for $d \in \mathbb{Z}_{2N}^*$, the invertible residues modulo $2N$. The automorphism $\tau_d(\cdot)$ acts on the monomials for $d = 2^\ell + 1$ with $1 \le \ell \le \log N$. We note that $\tau_d(X^i) = X^i$ for $i = \frac{N}{2^\ell} \cdot e$ and $\tau_d(X^i) = -X^i$ for $i = \frac{N}{2^\ell} \cdot o$, where e (resp., o) represents an even (resp., odd) integer. In other words, the map $\widehat{a} \mapsto \widehat{a} + \tau_d(\widehat{a})$ doubles the coefficients $\widehat{a}[i]$ if $i = \frac{N}{2^\ell} \cdot e$, but zeroizes the coefficients $\widehat{a}[i]$ if $i = \frac{N}{2^\ell} \cdot o$ (More details can be found in [11]).

Given the automorphisms $\tau_{2^\ell+1}$ for $\ell \in \{1, \ldots, \log N\}$ as the bases, any automorphism $\tau \in \text{Gal}(K/\mathbb{Q})$ can be expressed as a combination of them:

$$\tau = \tau_{2^1+1}^{b_1} \circ \tau_{2^2+1}^{b_2} \circ \cdots \circ \tau_{2^{\log N}+1}^{b_{\log N}}, b_\ell \in \{0, 1\}.$$

In general, for any $0 \le s < t \le \log N$, each automorphism $\tau \in \text{Gal}(K_t/K_s)$ can be expressed as $\tau = \tau_{2^{s+1}+1}^{b_{s+1}} \circ \cdots \circ \tau_{2^t+1}^{b_t} |_{K_t}$. In particular, we define $\tau(\widehat{a}) := \tau_{2^{s+1}+1}^{b_{s+1}} \circ \cdots \circ \tau_{2^t+1}^{b_t}(\widehat{a})$ for $\widehat{a} \in K, \tau \in \text{Gal}(K_t/K_s)$.

2.4 RLWE-based Homomorphic Encryption

Homomorphic encryption schemes based on RLWE, e.g., BGV [7], BFV [6, 16] and CKKS [12], are defined over a ring R. The decisional RLWE assumption states that the distribution in the form of $(-\widehat{a} \cdot \widehat{s} + \widehat{e}, \widehat{a}) \in R_q^2$ is computationally indistinguishable from the uniform distribution in R_q^2, where $\widehat{a} \leftarrow_\$ R_q$ and $\widehat{e} \leftarrow_\$ \chi_\sigma$ (a discrete Gaussian distribution with standard deviation σ over R_q) and $\widehat{s} \in R$ is a random secret. In this work, we use BFV as the basic HE scheme. Let t, q be two positive integers with $t \ll q$, R_q^2, R_t denote the ciphertext space and plaintext space, and q, t are the ciphertext modulus and plaintext modulus, respectively. The scheme has a set of public parameters $\text{HE.pp} = \{t, q, N, \sigma\}$ where N is the ring dimension. We will adopt the following algorithms to design an MVM protocol:

- **KeyGen.** We adopt the symmetric-key BFV-HE scheme [33], where the encryption and decryption depend on the same secret key $\widehat{sk}$, sampled uniformly from R_3 (i.e., each coefficient of polynomial $\widehat{sk}$ is sampled from $\{-1, 0, 1\}$).
- **Encode.** We use two coefficient encoding algorithms, denoted by Ecd_1 and Ecd_2. Specifically, Ecd_1 maps a vector $\mathbf{a} \in \mathbb{Z}_t^\ell$ to a polynomial $\widehat{a} \in R_t$ with $\widehat{a}[j] = \mathbf{a}[j]$ for $j \in [N]$ (padding zero if $\ell < N$). $\text{Ecd}_2(\mathbf{a})$ is the reciprocal polynomial of $\text{Ecd}_1(\mathbf{a})$, i.e., $\text{Ecd}_2(\mathbf{a}) = \widehat{a}(X^{-1}) \in R_t$. We note that the constant term of

$\text{Ecd}_1(\mathbf{a}) \cdot \text{Ecd}_2(\mathbf{b}) = \widehat{a}(X) \cdot \widehat{b}(X^{-1})$ is the inner product of $\mathbf{a}$ and $\mathbf{b}$ for two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{Z}_t^\ell$.

- **Decode.** Given a plaintext polynomial $\widehat{m} \in R_t$, this algorithm outputs the corresponding coefficient vector $\mathbf{m} \in \mathbb{Z}_t^\ell$.
- **Encrypt.** Given a plaintext polynomial $\widehat{m} \in R_t$, the encryption algorithm outputs a ciphertext $ct = (\widehat{c}_0, \widehat{c}_1)$, such that $\widehat{c}_0 = -\widehat{a} \cdot \widehat{sk} + \widehat{e} + \lfloor \frac{q}{t} \cdot \widehat{m} \rceil \in R_q$ and $\widehat{c}_1 = \widehat{a} \in R_q$ where $\widehat{a} \leftarrow_\$ R_q$ and $\widehat{e} \leftarrow_\$ \chi_\sigma$. We use Enc to denote the encryption algorithm, and also denote by $[\![\widehat{m}]\!]$ the ciphertext on $\widehat{m}$.
- **Decrypt.** Given a ciphertext $ct = (\widehat{c}_0, \widehat{c}_1) \in R_q^2$, the decryption algorithm outputs $\widehat{m} = \lfloor \frac{t}{q}(\widehat{c}_0 + \widehat{c}_1 \cdot \widehat{sk}) \rceil \in R_t$.
- **HomMulPt.** Given a polynomial $\widehat{a} \in R_t$ and a ciphertext $ct \in R_q^2$ on a plaintext polynomial $\widehat{m} \in R_t$, this algorithm HomMulPt outputs a ciphertext decrypted to $\widehat{m} \cdot \widehat{a} \in R_t$. This plaintext-ciphertext multiplication will amplify the noise from η to $\eta \cdot \eta_{mul}$ where η_{mul} is the multiplicative noise growth factor, which depends on the infinity norm of $\widehat{a}$. We refer the reader to [6, 16] for the details of HomMulPt.
- **HomAut.** Given a ciphertext ct on a plaintext polynomial $\widehat{m}$ and an automorphism $\tau \in \text{Gal}(K/\mathbb{Q})$, this algorithm homomorphically computes a new ciphertext $ct' \leftarrow \text{HomAut}(ct, \tau)$, which is decrypted to $\tau(\widehat{m})$. The homomorphic automorphism (a.k.a., rotation) operation will introduce an additive noise η_{aut}, and amplifies the noise from η to $\eta + \eta_{aut}$. The details of HomAut can be found in [6, 16].
- **PackLWEs.** Given $2^\ell (\le N)$ ciphertexts $\{ct_i\}_{i \in [2^\ell]}$ where ct_i encrypts $\widehat{m}_i \in R_t$, this algorithm PackLWEs homomorphically merge the constant terms of them into a new ciphertext ct, decrypted to $\widehat{m} \in R_t$ with $\widehat{m}[(N/2^\ell) \cdot i] = 2^\ell \cdot \widehat{m}_i[0]$ for $i \in [2^\ell]$. The algorithm details can be found in [11].
- **Expand.** Given an integer $\ell \le \log N$ and a ciphertext ct on a plaintext polynomial $\widehat{m} = \sum_{i=0}^{2^\ell - 1} \widehat{m}[i] \cdot X^i \in R_t$, this algorithm homomorphically expands it to 2^ℓ ciphertexts $\{ct_i\}_{i \in [2^\ell]}$, where ct_i is decrypted to $2^\ell \cdot \widehat{m}[i] \in \mathbb{Z}_t$. We write it as $\{ct_i\}_{i \in [2^\ell]} \leftarrow \text{Expand}(ct, \ell)$. We refer the reader to [2, 3] for details.

Additionally, we will invoke an algorithm (denoted by HomAdd) for computing homomorphic additions (see [6, 16] for details). In general, for both PackLWEs and Expand, a single invocation requires $2^\ell - 1$ calls of HomAut, and the value 2^ℓ in either $\{2^\ell \cdot \widehat{m}_i[0]\}$ or $\{2^\ell \cdot \widehat{m}[i]\}$ could be removed via multiplying the input ciphertexts by its inverse $2^{-\ell}$.

2.5 Conversion between Additive SS and HE

We use additive secret sharing (SS) over a ring $\mathbb{Z}_t$. Concretely, for an element $x \in \mathbb{Z}_t$, an additive SS on x (denoted by $\langle x \rangle$) is generated by sampling two shares $\langle x \rangle_0$ and $\langle x \rangle_1$ from $\mathbb{Z}_t$, such that $\langle x \rangle_0 + \langle x \rangle_1 = x \mod t$. The conversion from additive secret sharings to HE ciphertexts (denoted by A2H) and reverse direction (denoted by H2A) are described as follows.

Conversion from SS to HE. Suppose that P_i for $i \in \{0, 1\}$ holds a vector of additive secret sharings $\langle \mathbf{x} \rangle_i$, where $\langle \mathbf{x} \rangle_0 + \langle \mathbf{x} \rangle_1 = \mathbf{x} \mod t$. To convert the secret values from SS to HE, for each $i \in \{0, 1\}$, P_i encodes its shares by $\widehat{x}_i \leftarrow \text{Ecd}_1(\langle \mathbf{x} \rangle_i)$, and encrypts the resulting polynomial as $\text{Enc}(\widehat{x}_i)$. Then, P_i sends $\text{Enc}(\widehat{x}_i)$ to P_{1-i},

who performs the homomorphic addition $\text{Enc}(\widehat{x}_i) + \text{Ecd}_1(\langle\mathbf{x}\rangle_{1-i})$ to get the HE ciphertext $\text{Enc}(\widehat{x})$ such that $\widehat{x} = \text{Ecd}_1(\mathbf{x})$.

Conversion from HE to SS. Following prior works [13, 27], for the BFV scheme [6, 16] with coefficient encoding, the H2A conversion executes as follows:

(1) Given a ciphertext $\text{Enc}(\widehat{x})$ where $\widehat{x} \in R_t$ encodes a vector $\mathbf{x} \in \mathbb{Z}_t^N$, for some $i \in \{0, 1\}$, P_i samples $\widehat{r} \leftarrow_{\$} R_q$, and then sends $\text{Enc}(\widehat{x}) - \widehat{r}$ to P_{1-i}. P_i computes $\langle\mathbf{x}\rangle_i := \lfloor (t/q) \cdot \mathbf{r} \rceil$ as its own shares, where $\mathbf{r} \in \mathbb{Z}_q^N$ is the coefficient vector of $\widehat{r}$.

(2) P_{1-i} decrypts the ciphertext $\text{Enc}(\widehat{x}) - \widehat{r}$ to a polynomial, and then transforms the polynomial to the coefficient vector as its own shares $\langle\mathbf{x}\rangle_{1-i}$.

Using the above H2A procedure, the noise flooding [17] is eliminated, if a one-bit error is allowed (that is the case for PPML applications). For NTT encoding, noise flooding needs to be used, which leads to a noise 2^λ larger than that in normal encryption for statistical security parameter λ. This blows up the ciphertext modulus q of NTT encoding to at least λ bits larger than that of coefficient encoding.

2.6 Threat Model

Our protocol in the two-party setting is secure in the presence of any semi-honest adversary, who follows the protocol specification but tries to learn more than allowed from the protocol transcripts. Security against semi-honest adversaries has been considered in most of PPML protocols like [5, 27, 32]. We always assume that the adversary is probabilistic polynomial time (PPT). For PPML applications, one party (i.e., client) holds an input data, while the other party (i.e., server) holds the model parameters. Through jointly running a secure two-party computation (2PC) protocol, the client will obtain an inference result on the input data, and the server obtains nothing. For security, a party cannot learn any information on private data of the other party beyond the inference results known only by the client.

3 MATRIX VECTOR MULTIPLICATION

In this section, we present a new matrix vector multiplication (MVM) protocol, called *Rhombus*, shown in Figure 1. In particular, Alice (resp., Bob) holds a vector $\mathbf{v}$ (resp., a matrix $\mathbf{W}$), and the homomorphic-computation result will be converted to additive secret sharings. The conversion can be realized by invoking the H2A protocol shown in Section 2.5. Below, we focus on the part of homomorphic computation for MVM, and present the *input-output packing* (Section 3.1) and the *split-point picking* (Section 3.2) techniques to reduce the computation complexity based on the naive approach described in Section 1. Moreover, we also give the column-major based approach for MVM in Section 3.3, which is more suitable for tall matrices with large dimensions.

3.1 Input-Output Packing

The naive approach requires $O(r)$ of HomMulPt and HomAut calls. However, it fails to fully exploit the packing property of polynomials when $c < N$, thereby resulting in inefficient computation. For the matrix $\mathbf{W}$ with $c < N$, it is natural to encode N/c rows of $\mathbf{W}$ into a single plaintext polynomial, and then compute the N/c

inner products by only one HomMulPt. In this way, only rc/N HomMulPt calls are required, and we call it *input-packing*. However, the PackLWEs algorithm cannot be directly applied because the inner products are located at indices that are multiples of c in the coefficient vector of a polynomial, instead of only constant terms.

PackRLWEs. To solve the problem, we extend the PackLWEs algorithm to PackRLWEs, and introduce an additional parameter h that indicates each ciphertext has 2^h coefficients to be merged. We present it in Algorithm 1 and give an example in Figure 3. In Proposition 3.1, we show the correctness of the PackRLWEs algorithm.

Algorithm 1: PackRLWEs($\{ct_j\}_{j \in [2^\ell]}, h$)

Input: Ciphertexts $\{ct_j\}_{j \in [2^\ell]}$ and an integer $h \in [\log N + 1]$ with $\ell + h \leq \log N$.
Output: A single ciphertext ct.

1: **if** $\ell = 0$ **then**
2:      **return** $ct \leftarrow ct_0$
3: **else**
4:      $ct_e \leftarrow \text{PackRLWEs}(\{ct_{2j}\}_{j \in [2^{\ell-1}]}, h)$
5:      $ct_o \leftarrow \text{PackRLWEs}(\{ct_{2j+1}\}_{j \in [2^{\ell-1}]}, h)$
6:      $ct \leftarrow (ct_e + X^{\frac{N}{2^{\ell+h}}} \cdot ct_o) + \text{HomAut}(ct_e - X^{\frac{N}{2^{\ell+h}}} \cdot ct_o, \tau_{2^{\ell+h}+1})$
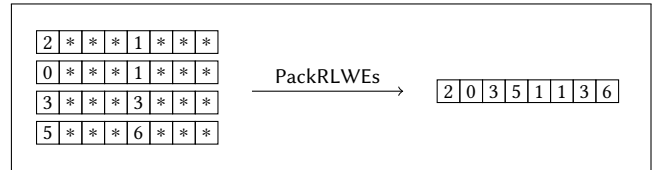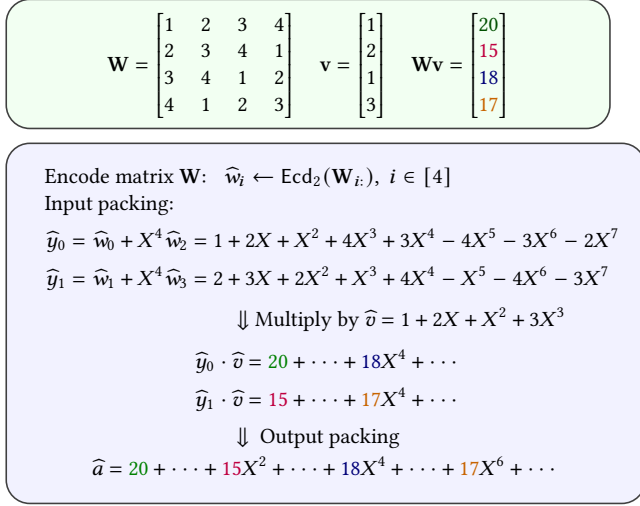7:   **return** ct
8: **end**

Figure 3: An example of PackRLWEs with $\ell = 2, h = 1, N = 8$

PROPOSITION 3.1. *Given ciphertexts $\{ct_j\}_{j \in [2^\ell]}$ (decrypted to $\widehat{a}_j$ for each $j \in [2^\ell]$) and an integer $h \in [\log N + 1]$ with $\ell + h \leq \log N$, PackRLWEs outputs a ciphertext ct which is decrypted to $\widehat{a}$, such that $\widehat{a}[(N/2^h) \cdot k + (N/2^{h+\ell}) \cdot j] = 2^\ell \cdot \widehat{a}_j[(N/2^h) \cdot k]$ for each $j \in [2^\ell], k \in [2^h]$.*

The correctness of the above proposition directly comes from [11]. The number of HomAut calls required by PackRLWEs with input 2^ℓ ciphertexts is still $2^\ell - 1$, which is independent of h, compared to that PackLWEs requires HomAut calls of $2^{\ell+h} - 1$. Note that PackRLWEs will degenerate to PackLWEs when $h = 0$.

With the extended packing algorithm PackRLWEs, one can realize the input-output packing to reduce the number of HomMulPt and HomAut calls to $O(rc/N)$. In particular, we refer to the final step of PackRLWEs as *output-packing*, which packs rc/N ciphertexts into one ciphertext. A toy example for the input-output packing process is shown in Figure 4 with $N = 8, r = 4$ and $c = 4$.

$$W = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{bmatrix} \quad v = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 3 \end{bmatrix} \quad Wv = \begin{bmatrix} 20 \\ 15 \\ 18 \\ 17 \end{bmatrix}$$

Encode matrix W: $\widehat{w}_i \leftarrow \text{Ecd}_2(W_{i:}),\ i \in [4]$
Input packing:
$\widehat{y}_0 = \widehat{w}_0 + X^4 \widehat{w}_2 = 1 + 2X + X^2 + 4X^3 + 3X^4 - 4X^5 - 3X^6 - 2X^7$
$\widehat{y}_1 = \widehat{w}_1 + X^4 \widehat{w}_3 = 2 + 3X + 2X^2 + X^3 + 4X^4 - X^5 - 4X^6 - 3X^7$

$\Downarrow$ Multiply by $\widehat{v} = 1 + 2X + X^2 + 3X^3$

$\widehat{y}_0 \cdot \widehat{v} = 20 + \cdots + 18X^4 + \cdots$
$\widehat{y}_1 \cdot \widehat{v} = 15 + \cdots + 17X^4 + \cdots$

$\Downarrow$ Output packing
$\widehat{a} = 20 + \cdots + 15X^2 + \cdots + 18X^4 + \cdots + 17X^6 + \cdots$

Figure 4: Input-output packing for $N = 8, r = c = 4$.

3.2 Split-Point Picking

The *input-output packing* described in Section 3.1 reduces the number of HomMulPt and HomAut calls to $O(rc/N)$. Accordingly, the experimental results in Figure 8 imply that the HomAut operation will dominate the whole computation. In the following, we present the critical technique called split-point picking (SPP), which further reduces the number of HomAut calls to $O(\sqrt{rc/N})$.

For the SPP technique, a key observation is that the input ciphertexts of PackRLWEs are obtained by multiplying different plaintexts with the same ciphertext (the encrypted vector) in the *input-output packing* algorithm. By separating these plaintexts from the ciphertexts, it becomes possible to apply HomAut solely on the common ciphertext.

We start with the *input-output packing* algorithm described as above, given the $r \times c$ matrix W and a ciphertext $[\![\widehat{v}]\!]$ as input, where W is encoded to $\{\widehat{y}_i\}_{i \in [m]}$ with $m = rc/N$ via the input packing. At this point, the ciphertexts input to PackRLWEs are in the form of $ct_i^0 = \widehat{y}_i \cdot [\![\widehat{v}]\!]$ for $i \in [m]$. Recall that PackRLWEs involves $\log m$ iterations. For the i-th ($1 \le i \le \log m$) iteration, it reduces the number of ciphertexts from $m/2^{i-1}$ to $m/2^i$, using the automorphism $\tau_{2^i(N/c)+1}$. We denote the $m/2^{i-1}$ input ciphertexts in the i-th iteration as $\{ct_j^{i-1}\}_{j \in [m/2^{i-1}]}$. We first focus on the computation in the 1st iteration, which transforms the set of ciphertexts $\{ct_j^0\}_{j \in [m]}$ into $\{ct_j^1\}_{j \in [m/2]}$, with $m/2$ HomAut calls. In particular,

$$ct_j^1 = (ct_{2j}^0 + X^{c/2} \cdot ct_{2j+1}^0) + \text{HomAut}(ct_{2j}^0 - X^{c/2} \cdot ct_{2j+1}^0, \tau_{2N/c+1})$$

for $j \in [m/2]$, according to step 6 in Algorithm 1. Based on the homomorphic property of $\tau_{2N/c+1}$, replacing ct_{2j}^0, ct_{2j+1}^0 with $\widehat{y}_{2j} \cdot [\![\widehat{v}]\!], \widehat{y}_{2j+1} \cdot [\![\widehat{v}]\!]$ respectively, we obtain

$$ct_j^1 = \left(\widehat{y}_{2j} + X^{c/2} \cdot \widehat{y}_{2j+1} \right) \cdot [\![\widehat{v}]\!] + \left(\tau_{\frac{2N}{c}+1} \left(\widehat{y}_{2j} \right) + X^{c/2} \cdot \tau_{\frac{2N}{c}+1} \left(\widehat{y}_{2j+1} \right) \right) \cdot \text{HomAut} \left([\![\widehat{v}]\!], \tau_{\frac{2N}{c}+1} \right) \quad (1)$$

From the above equation (1), we find that all HomAut operations are performed on the same ciphertext $[\![\widehat{v}]\!]$. Therefore, HomAut now

needs to be called only once in the 1st iteration. Note that the automorphisms on plaintexts are much faster than those on ciphertexts, hence, we ignore the automorphisms acting on plaintexts temporarily. Similarly, we can proceed to carry out the replacement for the 2nd iteration if $m \ge 4$. Specifically, the equation in this iteration is shown as follows:

$$ct_j^2 = \left(ct_{2j}^1 + X^{c/4} \cdot ct_{2j+1}^1 \right) + \text{HomAut} \left(ct_{2j}^1 - X^{c/4} \cdot ct_{2j+1}^1, \tau_{\frac{4N}{c}+1} \right)$$

for $j \in [m/4]$. Replacing ct_{2j}^1 with the ciphertext defined in the equation (1), we can derive the following equation:

$$ct_j^2 = \sum_{\tau \in \text{Gal}(K_{h+2}/K_h)} \left(\sum_{k=0}^3 \tau \left(\widehat{y}_{4j+k} \right) X^{\frac{kc}{4}} \right) \text{HomAut} \left([\![\widehat{v}]\!], \tau \right) \quad (2)$$

where $h = \log(N/c)$, $\text{Gal}(K_{h+2}/K_h) = \{1, \tau_{2^{h+1}+1}, \tau_{2^{h+2}+1}, \tau_{2^{h+2}+1} \circ \tau_{2^{h+1}+1}\}$. The equation (2) indicates that computing ct_j^2 in this manner requires only $\#\text{Gal}(K_{h+2}/K_h) - 1 = 3$ HomAut calls. From the replacements (1) and (2), we can summarize that $\#\text{Gal}(K_{h+L}/K_h) - 1 = 2^L - 1$ HomAut calls are required when computing ct_j^L for $j \in [m/2^L]$ in the replacement way. In the extreme case, $m - 1$ HomAuts are needed if we compute $ct_0^{\log m}$ using the above approach, which is equivalent to invoking the PackRLWEs directly.

Based on the derivation above, we will achieve the optimal efficiency, if switching the computation algorithm in the middle layer of the PackRLWEs tree. In particular, we adopt the replacement mentioned as above in the first $\lfloor (\log m)/2 \rfloor$ iterations of the PackRLWEs tree with HomAut calls of $2^{\lfloor (\log m)/2 \rfloor} - 1$. Note that the number of ciphertexts is $m/2^{\lfloor (\log m)/2 \rfloor}$ at this time. Then we invoke PackRLWEs on the input of $m/2^{\lfloor (\log m)/2 \rfloor}$ ciphertexts, which requires $m/2^{\lfloor (\log m)/2 \rfloor} - 1$ HomAut calls. Totally, the number of HomAut calls is $2^{\lfloor (\log m)/2 \rfloor} + m/2^{\lfloor (\log m)/2 \rfloor} - 2$, which amounts to a complexity of $O(\sqrt{m}) = O(\sqrt{rc/N})$. We refer to this replacement approach as *split-point picking*. In the context of MVM, we usually choose the split point corresponding to the middle layer in the PackRLWEs tree. We summarize the process in Algorithm 2.

PROPOSITION 3.2. *Given a matrix W with 2-power dimensions $r, c \le N$, an encrypted vector $[\![\mathbf{v}]\!]$ (encoding via Ecd_1) and the split-point u, Algorithm 2 outputs a ciphertext ct' which decrypts to $\widehat{a}'$, such that $\widehat{a}'[(N/r) \cdot i] = (W \cdot \mathbf{v})[i]$ for $i \in [r]$.*

The proposition 3.2 shows the correctness of Algorithm 2, and we defer the proof to the full version. We note that only the coefficients with indices being multiples of N/r are useful for the output ciphertext. Therefore, for the output ciphertext in the form of $(\widehat{c}_0, \widehat{c}_1) \in R_q^2$, we can drop the irrelevant coefficients of $\widehat{c}_0$ to reduce the communication overhead.

This work follows the MVM protocol framework shown in Figure 1, and only improves the part of locally homomorphic computation by invoking Algorithm 2. Therefore, the improved MVM protocol with Algorithm 2 is natural to be secure in the presence of semi-honest adversaries under the RLWE assumption, by directly following the security proof of previous protocols, e.g., [27].

3.3 Column Major based MVM

Section 3.2 gives the MVM algorithm by regarding the MVM as inner products between the vector $\mathbf{v}$ and each row of the matrix W.

Algorithm 2: RhombusMVM$_{RM}$: Row major based matrix vector multiplication with SPP optimization.

Input: $r \times c$ Matrix with $r, c \leq N$ (powers of two), a ciphertext $ct \leftarrow [\![\mathbf{v}]\!] \in R_q^2$, and the split-point u.
Output: A ciphertext ct'.

1: **for** $i \in [r]$ **do**
2: $\quad \widehat{w}_i \leftarrow \text{Ecd}_2(\mathbf{W}_{i:})$
3: **end**
4: $m \leftarrow r \cdot c/N, h \leftarrow \log(N/c)$
5: $ct \leftarrow m^{-1} \cdot ct$ // To remove the PackRLWEs factor
6: **for** $i \in [m]$ **do**
7: $\quad \widehat{y}_i \leftarrow \sum_{j=0}^{N/c-1} X^{j \cdot c} \cdot \widehat{w}_{i+j \cdot m}$ // Input packing
8: **end**
9: **for** $\tau \in \text{Gal}(K_u/K_h)$ **do**
10: $\quad ct_\tau \leftarrow \text{HomAut}(ct, \tau)$ // SPP: Replacement
11: **end**
12: **for** $i_1 \in [r/2^u]$ **do**
13: $\quad ct_{i_1} \leftarrow \text{Enc}(0)$
14: $\quad$ **for** $\tau \in \text{Gal}(K_u/K_h)$ **do**
15: $\quad\quad \widehat{z}_{i_1,\tau} \leftarrow \sum_{i_0=0}^{m2^u/r-1} X^{(N/2^u) \cdot i_0} \cdot \tau(\widehat{y}_{(r/2^u) \cdot i_0 + i_1})$
16: $\quad\quad ct_{i_1} \leftarrow \text{HomAdd}(ct_{i_1}, \text{HomMulPt}(ct_\tau, \widehat{z}_{i_1,\tau}))$
17: $\quad$ **end**
18: **end**
19: $ct' \leftarrow \text{PackRLWEs}(\{ct_{i_1}\}_{i_1 \in [r/2^u]}, u)$ // SPP: PackRLWEs
20: **return** ct'

Alternatively, MVM can also be viewed as the linear combination of the columns of the matrix with the vector. In this way, we can leverage the Expand algorithm described in Section 2.4 to compute MVM as follows.

(1) On input of a matrix $\mathbf{W}$ of dimension $r \times c$ and an encrypted vector $[\![\mathbf{v}]\!]$ with Ecd_1 for encoding, run the Expand algorithm to expand $[\![\mathbf{v}]\!]$ to c ciphertexts, each of them encrypts the value $\mathbf{v}[i]$ for $i \in [c]$.

(2) For $i \in [c]$, encode the column vector $\mathbf{W}_{:i}$ of the matrix $\mathbf{W}$ by Ecd_1 into a plaintext polynomial $\widehat{w}_i$.

(3) Execute the HomMulPt algorithm on these expanded ciphertexts and the encoded columns, and then accumulate these ciphertexts.

The correctness is straightforward based on the definition of Expand. The column-major approach requires c HomMulPt calls for computing the linear combination, and $c - 1$ HomAut calls for running Expand. Similar to the row-major approach, the number of HomAut calls can be further reduced to $O(\sqrt{c})$ using the SPP optimization. We provide the details of the column-major MVM algorithm in the full version.

It is easy to see that the row-major approach outperforms the column-major approach when both the number of rows and columns of the matrix are smaller than N because the latter cannot benefit from the *input packing*. More precisely, we can pack multiple rows of the matrix to one plaintext polynomial with row-major approach, but for column-major approach we cannot pack multiple columns

to one polynomial since each column will be multiplied by different element of the vector.

However, for matrix of large dimensions, i.e., the number of rows or columns is larger than N, the two approaches have their own merits. To analyze the case of large matrix, we suppose that $r = r'N, c = c'N$ for simplicity, where r', c' are positive integers. In this case, we have to partition the large matrix $\mathbf{W}$ into $r' \times c'$ submatrices, each of dimension $N \times N$, and partition the vector $\mathbf{v}$ into c' subvectors correspondingly, then perform the MVM on each pair of the submatrices and subvectors, respectively. We count the number of HomMulPt, HomAut calls and the complexity of matrix processing to compare the concrete efficiency of the two approaches. Before conducting the concrete comparison, we first observe the following facts:

- Both the two approaches require rc/N times of HomMulPt.
- With the row major approach, the PackRLWEs (Step 19 in Algorithm 2) can be called only once for the MVM computation with respect to the submatrices in the same row (note that the results of these MVMs will be accumulated), because it is additive homomorphic on packing two groups of ciphertexts: PackRLWEs($\{ct_i\}, u$) + PackRLWEs($\{ct_i'\}, u$) = PackRLWEs($\{ct_i + ct_i'\}, u$).
- With the column major approach, the similar fact holds true, i.e., the HomAut can also be merged, and thus be called only once, for the MVM computation with respect to the submatrices in the same row.
- For the matrix processing, larger split point u leads to more complex processing in both of the two algorithms.

Based on the analyses above, we can derive the concrete number of HomAut calls when employing the two approaches on large matrix (note that the *input packing* is not applicable), respectively. For row major approach with split point $u_r (0 \leq u_r \leq \log N)$, the vector $\mathbf{v}$ is encrypted to c' ciphertexts, each of them will go through $2^{u_r} - 1$ of HomAut calls (note that $h = 0$ in this case), in addition, for the MVMs with respect to the submatrices in the same row, we need $N/2^{u_r} - 1$ of HomAut calls inside the PackRLWEs. Totally, $c'(2^{u_r} - 1) + r'(N/2^{u_r} - 1)$ of HomAut are needed. For the column major approach with split point $u_c (0 \leq u_c \leq \log N)$, each of the c' input ciphertexts will go through the $N/2^{u_c} - 1$ of HomAut calls, besides, $2^{u_c} - 1$ of HomAut are needed for the MVMs with respect to the submatrices in the same row. Totally, $c'(N/2^{u_c} - 1) + r'(2^{u_c} - 1)$ of HomAut are needed when using column-major approach. As a result, the column major approach has the same complexity of HomAut calls as the row major approach if we set $u_c = \log N - u_r$, and the only difference lies on the matrix processing.
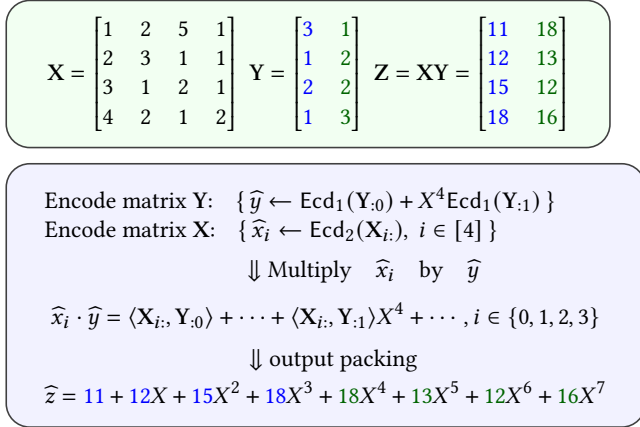
In summary, for the tall matrix with $r' > c'$ (i.e., $r > c$), the column major approach will outperform the row major approach. The reason is in this case, row major approach will need a larger $u_r (> (\log N)/2)$ to minimize the HomAut calls, leading to a more complex matrix processing than column-major one. However, for rectangular matrix with $r' < c'$ (i.e., $r < c$), row major will perform better than column major approach, due to a cheaper matrix processing in this case. We present the concrete efficiency comparison between these two approaches in Table 2 in Section 5.1.

4 MATRIX MULTIPLICATION

In this section, we focus on secure matrix multiplication, where Alice holds a matrix $Y_{m \times k}$ (to be encrypted), and Bob holds another matrix $X_{n \times m}$. Both parties will execute a protocol similar to Figure 1 to securely compute additive secret sharings of $X \cdot Y$. Although matrix multiplication could be naturally generalized from MVM, the cost of matrix multiplication with parameters (n, m, k) becomes k times that of MVM with matrix dimension (n, m) in this case. Especially when $n, m \ll N$, it results in significant communication waste. Below, we present two new matrix encoding methods that fully utilize all coefficients of the polynomial to achieve optimal asymptotic communication complexity. We consider the row-major approach in Section 4.1, and then show the corresponding column-major method in Section 4.2.

4.1 Matrix Encoding for Row-Major Approach

Recall that in the MVM algorithm shown in Section 3, the matrix is encoded in a packed manner, while the vector is encoded alone. When it comes to matrix multiplication, we can reverse the encoding manner of the matrix and vector to give a communication efficient method. That is, for the matrix Y to be encrypted, we pack multiple columns into one plaintext polynomial, similar to the encoding of matrices in MVM. Once the columns of matrix Y are packed, each row of matrix X must be encoded separately, and then has to be multiplied with the encrypted Y. Figure 5 gives an example for the case that $n = m, mk = N$ where n, m, k are powers of two. Furthermore, we can likewise apply the SPP approach. We describe it in Algorithm 3.

$$X = \begin{bmatrix} 1 & 2 & 5 & 1 \\ 2 & 3 & 1 & 1 \\ 3 & 1 & 2 & 1 \\ 4 & 2 & 1 & 2 \end{bmatrix} \quad Y = \begin{bmatrix} 3 & 1 \\ 1 & 2 \\ 2 & 2 \\ 1 & 3 \end{bmatrix} \quad Z = XY = \begin{bmatrix} 11 & 18 \\ 12 & 13 \\ 15 & 12 \\ 18 & 16 \end{bmatrix}$$

Encode matrix Y: $\{ \widehat{y} \leftarrow \text{Ecd}_1(Y_{:0}) + X^4 \text{Ecd}_1(Y_{:1}) \}$
Encode matrix X: $\{ \widehat{x_i} \leftarrow \text{Ecd}_2(X_{i:}), \ i \in [4] \}$

$$\Downarrow \text{Multiply} \quad \widehat{x_i} \quad \text{by} \quad \widehat{y}$$

$$\widehat{x_i} \cdot \widehat{y} = \langle X_{i:}, Y_{:0} \rangle + \cdots + \langle X_{i:}, Y_{:1} \rangle X^4 + \cdots, \ i \in \{0, 1, 2, 3\}$$

$$\Downarrow \text{output packing}$$

$$\widehat{z} = 11 + 12X + 15X^2 + 18X^3 + 18X^4 + 13X^5 + 12X^6 + 16X^7$$

Figure 5: Example for row-major based matrix multiplication with $n = m = 4, k = 2, N = 8$.

For general dimensions n, m, k of matrices, we have to partition them into several blocks satisfying the above requirements, and perform the computation on each pair of blocks of X and Y using Algorithm 3. In the following, we give two partition methods V1, V2, which enable a trade-off between the matrix encoding and homomorphic automorphism.

Matrix Partition V1. The first method is to partition the matrix Y vertically. Specifically, we pad m to a power of 2, denoted by $\overline{m}$. Then, every $N/\overline{m}$ adjacent columns of Y will be packed into a single

---

Algorithm 3: RhombusMatMul$_{RM}$: Row major based matrix multiplication with the SPP optimization.

Input: Matrix $X_{n \times m}$, encrypted matrix $ct \leftarrow [\![Y_{m \times k}]\!]$ with $n = m, mk = N$, and the split-point u.

**Output:** $ct' \leftarrow [\![X \cdot Y]\!]$

1: $ct \leftarrow m^{-1} \cdot ct$ // Remove the PackRLWEs factor.

2: $h \leftarrow \log k$

3: **for** $i \in [n]$ **do**

4: $\quad \widehat{x_i} \leftarrow \text{Ecd}_2(X_{i:})$ // Encode each row of X separately.

5: **end**

6: **for** $\tau \in \text{Gal}(K_u/K_h)$ **do**

7: $\quad ct_\tau \leftarrow \text{HomAut}(ct, \tau)$ // SPP: Replacement

8: **end**

9: **for** $i_1 \in [nk/2^u]$ **do**

10: $\quad ct_{i_1} \leftarrow \text{Enc}(0)$

11: $\quad$ **for** $\tau \in \text{Gal}(K_u/K_h)$ **do**

12: $\quad\quad \widehat{z}_{i_1, \tau} \leftarrow \sum_{i_0 = 0}^{2^{u-h}-1} X^{(N/2^u) \cdot i_0} \cdot \tau(\widehat{x}_{i_0 \cdot (nk/2^u) + i_1})$;

13: $\quad\quad ct_{i_1} \leftarrow \text{HomAdd}(ct_{i_1}, \text{HomMulPt}(ct_\tau, \widehat{z}_{i_1, \tau}))$

14: $\quad$ **end**

15: **end**

16: $ct' \leftarrow \text{PackRLWEs}(\{ct_{i_1}\}_{i_1 \in [nk/2^u]}, u)$ // SPP: PackRLWEs

17: **return** ct'

---

polynomial, and thus Y will be encrypted to $\lceil k\overline{m}/N \rceil$ ciphertexts. For the matrix X, we divide it into $\lceil n/\overline{m} \rceil$ blocks horizontally, each of them is of dimension $\overline{m} \times \overline{m}$ (padding zero for the last block if necessary). In this way, each block of X will be multiplied by every ciphertext on Y, as shown in the left part of Figure 6.

Matrix Partition V2. Instead of padding m, we can also pad k to a power of 2, denoted by $\overline{k}$. Then, we divide the matrix Y into $\lceil m\overline{k}/N \rceil$ blocks horizontally, each of dimension $N/\overline{k} \times \overline{k}$. Accordingly, X is partitioned into $\lceil n\overline{k}/N \rceil \times \lceil m\overline{k}/N \rceil$ blocks, where each block is of dimension $N/\overline{k} \times N/\overline{k}$, as shown in the right part of Figure 6.
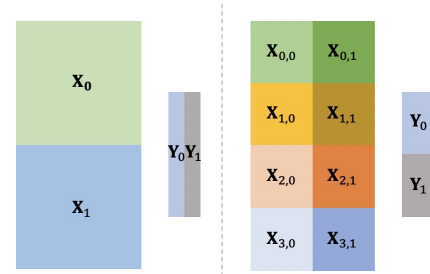
Figure 6: Two methods of partitioning matrices X, Y

Below, we analyze the following performance characteristics for the two partition methods with SPP optimization:

V1. The number of HomMulPt calls is $n \cdot \lceil \overline{m}k/N \rceil$. The total number of HomAut calls is $\lceil \overline{m}k/N \rceil (s_0 - 1) + \lceil \overline{m}k/N \rceil \lceil n/\overline{m} \rceil (s_1 - 1)$ where $s_0 s_1 = \overline{m}$ implicitly gives the split point. The total number of ciphertexts sent between the two parties is $\lceil \overline{m}k/N \rceil + \lceil n/\overline{m} \rceil \lceil \overline{m}k/N \rceil$.

V2. The number of HomMulPt calls is $n \cdot \lceil m\bar{k}/N \rceil$. For HomAut, we note that the HomAut operations for the blocks of $\mathbf{X}$ in the same row could be merged and performed only once. As a result, the number of HomAut operations is $\lceil m\bar{k}/N \rceil(s_0 - 1) + \lceil n\bar{k}/N \rceil(s_1 - 1)$, where $s_0 s_1 = N/\bar{k}$ indicates the split point. The number of ciphertexts sent is $\lceil m\bar{k}/N \rceil + \lceil n\bar{k}/N \rceil$.

Table 3 gives the comparison with prior works for matrix multiplication, where Encode/NTT represents the number of plaintexts of the encoded matrix $\mathbf{X}$. We note that the matrix $\mathbf{X}$ will be encoded to multiple plaintexts, and then these plaintexts will be transformed to the NTT form and are multiplied by the encrypted blocks of $\mathbf{Y}$. The encoding and transformation (to NTT) could be performed only once. The number of HomAut calls in Table 3 are the minimized by choosing the optimal split-point.

4.2　Matrix Encoding for Column-Major Method

Similar to the column-major MVM described in Section 3.3, matrix multiplication can also be realized with the Expand algorithm. Specifically, we note that $\mathbf{X} \cdot \mathbf{Y} = \sum_{i \in [m]} \mathbf{X}_{:i} \cdot \mathbf{Y}_{i:}$. Taking the product of $\mathbf{X}_{:0}$ and $\mathbf{Y}_{0:}$ as an example, this tensor product can be computed as shown in Figure 7 (where $\mathbf{X}, \mathbf{Y}$ are the same as that in Figure 5).

Encode $\mathbf{Y}_{0:}$:　$\widehat{y}_0 = 3 + 1X^4$
Encode $\mathbf{X}_{:0}$:　$\widehat{x}_0 = 1 + 2X + 3X^2 + 4X^3$

$$\Downarrow \text{Multiply}　\widehat{x}_0 \quad \text{and} \quad \widehat{y}_0$$

$$\widehat{x}_0 \cdot \widehat{y}_0 = 3 + 6X + 9X^2 + 12X^3 + 1X^4 + 2X^5 + 3X^6 + 7X^7$$

Figure 7: Tensor of $\mathbf{X}_{:0}$ and $\mathbf{Y}_{0:}$ from HE with $N = 8$.

Note that $\mathbf{Y}$ is encrypted in a packed manner. Hence, our goal is to extract the $\widehat{y}_0, \widehat{y}_1, \widehat{y}_2, \widehat{y}_3$, in the same encoding manner as in Figure 7. Exactly, the extraction can be realized from the Expand algorithm by setting the parameter $\ell = 2$.

For the general matrix parameters n, m, k, we can follow the same matrix partition methods described in Section 4.1. In this way, the column-major matrix multiplication has the same complexity as the row-major approach in terms of HomAut, HomMulPt and Encode/NTT operations. The only difference lies in the processing of plaintext matrices (similar to MVM described in Section 3.3). In particular, for the dimension (n, m) of the matrix $\mathbf{X}$, the column major approach has the cheaper matrix processing than the row major approach when $n > m$.

5　PERFORMANCE EVALUATION

In this section, we evaluate the performance of *Rhombus*, and compare it with prior arts. This includes the efficiency comparison of MVM, matrix multiplication, pointwise convolution and the end-to-end inference of ResNet50. We adopt the SEAL library [53] of version 3.7 with HEXL acceleration to implement HE-based MVM and matrix multiplication. For the pointwise convolution and ResNet50 inference, our implementation builds upon the open-sourced PPML library OpenCheetah [26], and simulates the network environment with EMP-toolkit [55].

Experimental Setup. All our experiments were conducted using Intel(R) Xeon(R) Platinum 8358P CPU at 2.60GHz with 32GB of memory. We run our benchmarks in two network settings: LAN (3Gbps of bandwidth and 0.3ms of round-trip time) and WAN (100Mbps of bandwidth and 40ms of round-trip time). All experiments are run with 4 threads. We use the tc tool on Linux to control the network traffic.

HE parameters. Our protocols select the dimension $N = 8192$ of polynomial ring R, ciphertext modulus $q \approx 2^{100}$, plaintext modulus $t = 2^{37}$ and standard deviation $\sigma = 3.2$ of a discrete Gaussian distribution, which achieves 128-bit security level. For two recent protocols [5, 27] to be compared, *Cheetah* chooses the HE parameters as HE.pp[*Cheetah*] $= \{N = 4096, q \approx 2^{105}, t = 2^{37}, \sigma = 3.2\}$, and *HELiKs* sets the HE parameters as HE.pp[*HELiKs*] $= \{N = 8192, q \approx 2^{120}, p \approx 2^{37}, \sigma = 3.2\}$, which enables the noise flooding to provide 32-bit statistical security according to the smudging lemma [4], where p is the plaintext modulus. In Figure 8, we provide the overhead (including the runtime and noise growth) of the basic homomorphic operations, which helps to understand the improvements we have made compared to previous approaches.
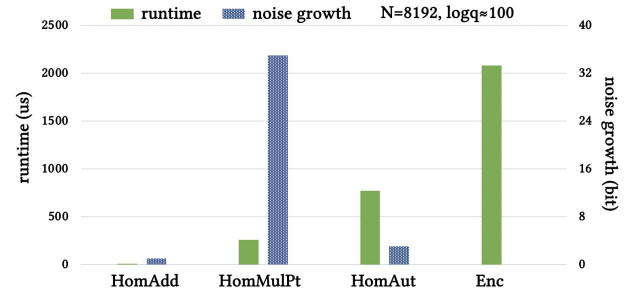
Figure 8: Running time (in μs) and noise growth (in bit) of homomorphic operations for parameters $N = 8192$ and $\log q \approx 100$.

5.1　Matrix Vector Multiplication

In Table 1, we give a theoretical comparison between our MVM protocol *Rhombus* and previous protocols, where HomAut, HomMulPt and Encrypt are three dominated operations. We note that for the NTT encoding based method, such as *HELiKs* [5], the multiplicative noise $\widehat{\eta}_{mul}$ introduced by HomMulPt will be greater than (or equal to) η_{mul} of coefficient encoding. This is because the noise growth factor depends only on the value of the coefficients of the plaintext polynomial and not on the size of the plaintext modulus, and in most cases, the NTT encoding will transform a vector even with small elements to a polynomial whose coefficients are spread throughout the entire plaintext space.

We use Algorithm 2 and choose the optimal split-point to minimize the number of rotations (i.e., HomAut) called by *Rhombus*. We focus on analyzing the comparison between *Rhombus* and the previous approaches listed in Table 1. For the approaches with NTT encoding, *HELiKs* has the lower complexity than *GAZELLE* and *HS*, and has a smaller noise growth compared to *GALA* by swapping the order of HomMulPt and Rotation, such that *HELiKs* can choose a smaller ciphertext modulus q than *GALA*. For the coefficient encoded approaches, we note that the $O(r)$ number of HomAut calls in *CHAM* is more costly than the $O(\sqrt{rc/N})$ number of Encrypt

operations in *Cheetah* in most cases according to Figure 8, and *CHAM* adopted the FPGA hardware to accelerate their approach to achieve a comparable performance as *Cheetah*. Therefore, in terms of concrete performance (shown below), we only conduct the comparsion with the two state-of-the-art protocols, i.e., *HELiKs* with NTT encoding (over a field $\mathbb{F}_p$) and *Cheetah* with coefficient encoding (over a ring $\mathbb{Z}_t$). Compared to *Cheetah*, *Rhombus* (Algorithm 2) requires additional $O(\sqrt{rc/N})$ rotations. However, *Cheetah* requires additional $O(\sqrt{rc/N})$ calls of Encrypt (and decryption), which is more expensive than rotation, as shown in Figure 8. For communication, *Cheetah* incurs a significant overhead, especially for matrices with large dimensions. While *HELiKs* requires the number of rotations linear in the matrix dimension, *Rhombus* needs only *sublinear* number of rotations. Furthermore, *Rhombus* can drop the unused coefficients of ciphertext polynomials to further reduce the communication cost, which is *not* supported by *HELiKs* with NTT encoding.

For concrete performance (running time and concrete communication cost), we compare *Rhombus* with the state-of-the-art protocols *HELiKs* and *Cheetah*, which is reported in Figure 9. In the WAN setting, we observe that *Rhombus* achieves a speedup of up to 8× over *HELiKs* and 3.4× over *Cheetah*. In the LAN setting, we find that *Rhombus* improves the performance of *Cheetah* by 1.8× and that of *HELiKs* by 7.4×. In terms of communication cost, *Rhombus* reduces the communication by a factor of 21× compared to *Cheetah*, and achieves the improvement of 1.8× over *HELiKs*.

In Table 2, we present the comparison between row major approach and column major approach on large matrices of different dimensions. As analyzed in 3.3, we choose the optimal split point u to minimize the number of HomAut calls, such that the only differences between the two approaches lies in the matrix processing. In particular, when $r < c$, the row major approach has a cheaper matrix processing than the column major approach, and thus performs better than the column major approach, conversely, the column major approach outperforms the row major approach when $r > c$. The time difference between the two approaches come from the different complexity of the matrix processing. And the two approaches perform comparably when $r = c$.

5.2 Matrix Multiplication

Firstly, we show a theoretical comparison between the secure matrix-multiplication protocol *Rhombus* (described in Section 4) and the most efficient protocols in the two-party setting, which is shown in Table 3. Compared to *Iron*, our protocol *Rhombus* takes about $O(\sqrt{N/k})$ times less communication, where $k \ll N$ for the parameters used in real-world applications. As a trade-off, *Rhombus* requires additional rotations of either $O(\frac{mk\sqrt{n}}{N})$ or $O(\sqrt{\frac{nmk}{N}})$, which is *sublinear* in the dimension (either n or m) of a matrix. *BumbleBee* [38] proposed the IntrLeave algorithm to realize the same functionality as our PackRLWEs, but required $O(\min\{\ell \cdot 2^\ell, N\})$ rotation calls when packing 2^ℓ ciphertexts, with each having $N/2^\ell$ coefficients to be merged, where PackRLWEs only requires $O(2^\ell)$ rotations and $\ell \leq \log N$. While *BumbleBee* requires the almost linear complexity for the number of rotations, *Rhombus* adopts the SPP optimization, which further reduces the

number of rotations to sublinear complexity. Therefore, *Rhombus* has a significantly lower computation cost than *BumbleBee*. Protocol *BOLT* [43] adopts NTT encoding, and requires a noise growth larger than *Rhombus*, where $\widehat{\eta}_{mul} \geq \eta_{mul}$. For the parameters N, n, m, k used in real-world applications, *BOLT* requires significantly more either encoding operations than *Rhombus-V1* or rotation operations than *Rhombus-V2*. Therefore, *Rhombus* has the lower computation cost than *BOLT*. Furthermore, *BOLT* with NTT encoding not only incurs more complex computation for encoding matrices, but also requires additional overhead when converting the computational results to additive secret sharings over a ring (used by *BOLT* to compute non-linear layers).

For the end-to-end performance of two-party matrix multiplication, we compare *Rhombus* with the recent protocols *Iron* [21] and *BumbleBee* [38], reported in Table 4. The HE parameters for *Iron* and *BumbleBee* are set as HE.pp[*Iron*] = $\{N = 4096, q \approx 2^{92}, t = 2^{37}, \sigma = 3.2\}$ and HE.pp[*BumbleBee*] = $\{N = 8192, q \approx 2^{112}, t = 2^{37}, \sigma = 3.2\}$, respectively. Note that the code of *BOLT* is not publicly available for now, and thus we only conduct a theoretical comparison with *BOLT*. For different matrix dimensions, we choose a suitable matrix partition method (V1 or V2) described in Section 4.1 to achieve a better performance. Compared to *Iron*, *Rhombus* achieves a speedup by 2.7× ∼ 3.9× (resp., 1.3× ∼ 2.3×) in the WAN (resp., LAN) setting, and reduces the communication cost by a factor of 4.6× ∼ 8×. Compared to *BumbleBee*, *Rhombus* demonstrates a performance improvement by a factor of 1.9× ∼ 3.9× (resp. 2.5× ∼ 4.5×) in the WAN (resp., LAN) setting, and achieves a communication reduction by around 1.1×.

5.3 Pointwise Convolutions

3D convolution is a common linear operation used in DNNs. In most cases, the convolutional kernels/filters are quite small, such as 3×3 and 1×1. In particular, pointwise convolution (with filter size 1×1) is commonly used to combine feature channels, reduce dimensionality and decrease the model's parameters. For example, in SqueezeNet [29] and DenseNet [24], half of the convolutions are pointwise; in ResNet50, ResNet101 and ResNet152 [22], there are about two-thirds of the convolutions are pointwise. Although *Cheetah* has proposed a generic method for convolution computation, it performs poorly for pointwise convolutions. Specifically, for the inference of ResNet50, the overhead for pointwise convolutions already accounts for 80% of all convolution layers.

As an application of our matrix-multiplication protocol *Rhombus*, we implemented the pointwise convolution and evaluate its performance. In particular, pointwise convolution of a (c_i, h, w) input image with $(c_o, f_h = 1, f_w = 1)$ filter of stride s could be transformed to the matrix multiplication with parameters $(hw/s^2, c_i, c_o)$ [5] without any additional costs. We report the performance comparison in our full version under the WAN setting among *Rhombus*, *Cheetah* [27] and *HELiKs* [5] for pointwise convolutions in ResNet50. The experiments show that *HELiKs* (resp., *Cheetah*) takes 338 MB (resp., 842 MB) of total communication for all pointwise convolutions, and *Rhombus* needs the total communication of 182 MB, which achieves the improvement by 1.9× (resp. 4.6×), compared to *HELiKs* (resp., *Cheetah*). For running time of computing all pointwise convolutions, while *HELiKs* (resp., *Cheetah*) takes 56 seconds

Table 1: Comparison of HE-based MVM protocols between Rhombus and known protocols in the two-party setting, where $r, c \leq N$ are the number of rows and columns in a matrix respectively, c is also the dimension of a vector, and N is the dimension of polynomial ring R in HE. The computational costs of the basic operations satisfy: Encrypt $>$ Rotation $>$ HomMulPt, as shown in Figure 8. η is the initial noise in the input ciphertext, η_{mul} (resp., $\widehat{\eta}_{mul}$) is the noise introduced by HomMulPt with coefficient (resp., NTT) encoding, which satisfies $\widehat{\eta}_{mul} \geq \eta_{mul}$. η_{aut} is the noise introduced by HomAut.

Protocol	Rotation (i.e., HomAut)	HomMulPt	Encrypt& Comm.	Encoding	Noise growth
HS [19]	$O(\max\{\sqrt{r}, \sqrt{c}\})$	$O(\max\{r, c\})$	$O(1)$	NTT	$(\eta + \eta_{aut}) \cdot \widehat{\eta}_{mul} + \eta_{aut}$
GAZELLE [32]	$O(\frac{rc}{N} + \log(\frac{N}{c}))$	$O(\frac{rc}{N})$	$O(1)$	NTT	$(\eta + \eta_{aut}) \cdot \widehat{\eta}_{mul} + \eta_{aut}$
GALA [58]	$O(\frac{rc}{N})$	$O(\frac{rc}{N})$	$O(1)$	NTT	$(\eta + \eta_{aut}) \cdot \widehat{\eta}_{mul}$
HELiKs [5]	$O(\frac{rc}{N})$	$O(\frac{rc}{N})$	$O(1)$	NTT	$\eta \cdot \widehat{\eta}_{mul} + \eta_{aut}$
CHAM [49]	$O(r)$	$O(\frac{rc}{N})$	$O(1)$	Coefficient	$\eta \cdot \eta_{mul} + \eta_{aut}$
Cheetah [27]	0	$O(\frac{rc}{N})$	$O(\sqrt{\frac{rc}{N}})$	Coefficient	$\eta \cdot \eta_{mul}$
Rhombus	$O(\sqrt{\frac{rc}{N}})$	$O(\frac{rc}{N})$	$O(1)$	Coefficient	$(\eta + \eta_{aut}) \cdot \eta_{mul} + \eta_{aut}$
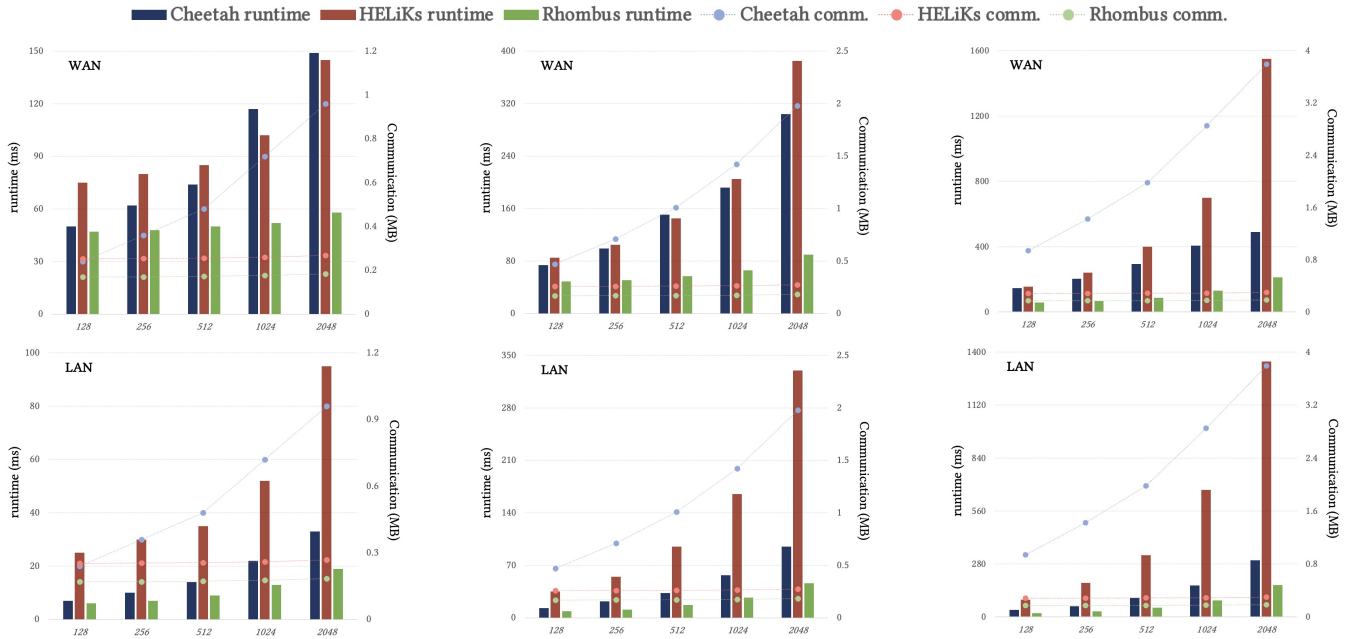
Figure 9: Comparison of the end-to-end performance of secure two-party MVM protocols for matrices with different dimensions in the LAN and WAN settings. Comparison of communication efficiency is also involved in this figure. For six sub-figures, the number of columns in a matrix are 256 (two sub-figures on the left), 1024 (two sub-figures in the middle) and 4096 (two sub-figures on the right) respectively, and these numbers are also the dimensions of vectors. The number of rows in a matrix is shown in the x-coordinate of these sub-figures, i.e, 128, 256, 512, 1024 and 2048.

Table 2: Comparison between row major and column major approaches for large matrices of different dimensions, the performance data includes only the time (measured in seconds) for the computation of MVM, excluding the time for network communication because the two approaches have the same communication overhead.

Approach	Dimension ($r \times c$)					
	$2^{13} \times 2^{14}$	$2^{14} \times 2^{13}$	$2^{14} \times 2^{14}$	$2^{13} \times 2^{15}$	$2^{15} \times 2^{13}$	$2^{15} \times 2^{15}$
row major, 3.2	1.32	1.68	2.66	2.56	3.86	10.8
column major, 3.3	1.68	1.33	2.67	3.88	2.55	10.8

Table 3: Complexity comparison of HE-based two-party matrix-multiplication protocols among Rhombus, Iron, BumbleBee and BOLT. Two matrices have dimensions $n \times m$ and $m \times k$, respectively. N denotes the polynomial-ring dimension in HE. *Rhombus-V1* (resp., *Rhombus-V2*) adopts the matrix partition method V1 (resp., V2) shown in Section 4.1. η is the initial noise of the input ciphertexts.

	Rotation (i.e., HomAut)	HomMulPt	Encode/NTT	Communication	Encoding	Noise growth
Iron [21]	0	$O(\frac{nmk}{N})$	$O(\frac{nmk}{N})$	$O(\sqrt{\frac{nmk}{N}})$	Coefficient	$\eta \cdot \eta_{mul}$
BumbleBee [38]	$O(\frac{kn \log N}{\sqrt{N}})$	$O(\frac{nmk}{N})$	$O(\frac{nm}{\sqrt{N}})$	$O(\frac{km}{N} + \frac{kn}{\sqrt{N}})$ †	Coefficient	$\eta \cdot \eta_{mul} + \eta_{aut}$
BOLT [43]	$O(\frac{mk\sqrt{n}}{N})$	$O(\frac{nmk}{N})$	$O(\frac{nmk}{N})$	$O(\frac{k(n+m)}{N})$	NTT	$(\eta + \eta_{aut}) \cdot \widehat{\eta}_{mul} + \eta_{aut}$
Rhombus-V1	$O(\frac{mk\sqrt{n}}{N})$	$O(\frac{nmk}{N})$	$O(n)$	$O(\frac{k(n+m)}{N})$	Coefficient	$(\eta + \eta_{aut}) \cdot \eta_{mul} + \eta_{aut}$
Rhombus-V2	$O(\sqrt{\frac{nmk}{N}})$	$O(\frac{nmk}{N})$	$O(\frac{nmk}{N})$	$O(\frac{k(n+m)}{N})$	Coefficient	$(\eta + \eta_{aut}) \cdot \eta_{mul} + \eta_{aut}$

† The complexity is derived from the partition window $n_w = 1, m_w = \sqrt{N}, k = \sqrt{N}$ given in *BumbleBee*.

Table 4: Comparison of the end-to-end performance and communication efficiency between Rhombus with Iron for secure matrix multiplication. The dimensions of matrices are chosen as same as that in the BERT-base model [15] with input of 128 tokens. We refer the reader to [21, 43] for more details.

Dimension (n, m, k)	Protocol	Running Time (s)		Comm. (MB)
		LAN	WAN	
(128, 128, 64)	*Iron* [21]	0.08	0.35	1.75
	BumbleBee [38]	0.16	0.23	0.24
	Rhombus-V1	0.06	0.12	0.22
(768, 768, 128)	*Iron* [21]	1.27	2.96	15.6
	BumbleBee [38]	2.39	2.45	2.93
	Rhombus-V2	0.63	0.75	2.7
(3072, 768, 128)	*Iron* [21]	4.86	7.67	31.3
	BumbleBee [38]	9.4	10	7.35
	Rhombus-V2	2.07	2.55	6.75
(768, 3072, 128)	*Iron* [21]	5.26	8.52	31.3
	BumbleBee [38]	7.2	7.4	7.3
	Rhombus-V2	2.88	3.12	6.75

(resp., 101 seconds), *Rhombus* runs in about 30 seconds, which gives a speedup of 1.9× (resp., 3.4×). We also measured the performance of pointwise convolutions in ResNet50 under the LAN setting for these protocols, which is given in the full version. In particular, while *HELiKs* (resp., *Cheetah*) takes 35 seconds (resp. 30 seconds) for all pointwise convolutions, *Rhombus* needs 15 seconds, which achieves the improvement by 2.3× (resp., 2×).

5.4 End-to-End PPML Inference

We integrate *Rhombus* into the OpenCheetah library to demonstrate the efficiency improvement of end-to-end PPML inference. We use the popular DNN model ResNet50 as an example just like as in *HELiKs* [5]. In Figure 10, we illustrate the comprehensive comparison of running time between *Rhombus* and the state-of-the-art protocols *Cheetah* [27] and *HELiKs* [5] for the end-to-end inference of ResNet50. In this figure, we divide the convolutions into two parts: pointwise convolutions (pw-conv) and non-pointwise

convolutions (npw-conv). Figure 11 reports the comparison of communication cost among these protocols for the end-to-end inference of ResNet50. The experiments are run under two frameworks: (1) one is *CryptFlow2* [48] (CF2 in short) where all sub-protocols builds over a field $\mathbb{F}_p$ with $p \approx 2^{37}$; (2) the other is *Cheetah* [27], in which all sub-protocols work over a ring $\mathbb{Z}_{2^{37}}$. Below, we give the performance analysis of ResNet50 inference.

Performance of HELiKs. The recent work [5] integrated *HELiKs* into the CF2 framework [48] by replacing the original matrix-multiplication and 3D-convolution protocols with their new protocols, which is called *HELiKs+CF2*. Additionally, *HELiKs* performed the encoding of convolutional filters and NTT transformation in the preprocessing phase to improve the performance of the online phase. In our benchmark, we focus on the end-to-end running time and total communication without dividing the protocol into the preprocessing and online phases. We note that both *Cheetah* and *Rhombus* can also benefit from executing a similar preprocessing. *HELiKs+CF2* works a field $\mathbb{F}_p$, which leads to be less efficient than the best-known protocol over a ring $\mathbb{Z}_t$ for secure computation of non-linear layers. Besides, *HELiKs+CF2* uses the IKNP OT extension [30] in the implementation, which incurs a large communication. The experimental result shows that *HELiKs+CF2* executes one inference in 42 minutes (resp., 195 seconds) in the WAN (resp., LAN) setting and the communication cost of 28.8 GB.

Performance of Cheetah. *Cheetah* [27] builds upon a ring $\mathbb{Z}_{2^{37}}$. For non-linear layers, *Cheetah* employs Ferret OT extension [56] with sublinear communication and uses it to realize the approximate truncation, which significantly reduces the communication cost of non-linear operators. For linear layers, *Cheetah* adopted the coefficient encoding to compute 3D convolution and matrix multiplication. However, *Cheetah* does not fully utilize each coefficient of the polynomial, resulting in significant waste in both communication and computation. The experimental result shows that *Cheetah* completes one inference within 213 seconds in the WAN setting and 58 seconds in the LAN setting. Figure 10 shows the proportion of running time of each module in *Cheetah*, among which 50% of running time takes at the pointwise convolutions. The total communication of *Cheetah* is 1.77 GB, while 46% of communication is used for computing pointwise convolutions.

Performance of Rhombus+Cheetah. We integrate *Rhombus* into the *Cheetah* framework by replacing the original pointwise convolution and MVM protocols with our protocols. Using our protocol *Rhombus*, the end-to-end running time for ResNet50 inference is improved to 141 seconds (resp., 42 seconds) in the WAN (resp., LAN) setting, and the total communication is reduced to 1.13 GB. In terms of the end-to-end performance, we achieve the improvement of 18× (resp., 1.5×) in the WAN setting and the improvement of 4.6× (resp., 1.4×) in the LAN setting, compared to *HELiKs* (resp., *Cheetah*). In terms of communication efficiency, *Rhombus* achieves the improvements by 26× over *HELiKs* and 1.6× over *Cheetah*, as shown in Figure 11. Due to the performance advantage of computing pointwise convolutions as shown in Section 5.3, our solution *Rhombus* will also achieve better efficiency for secure two-party inference of DNN models SqueezeNet, DenseNet, ResNet101 and ResNet152, compared to the state-of-the-art approaches *HELiKs* and *Cheetah*, where these DNNs have at least a half of convolutions are pointwise convolutions.
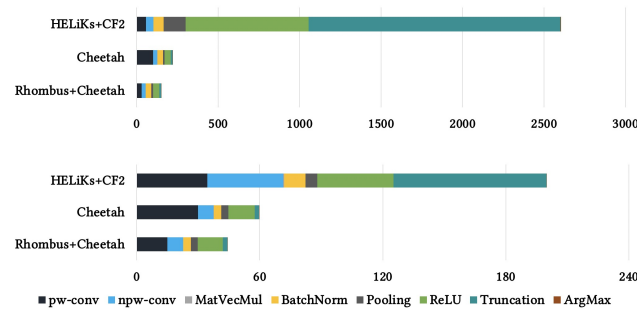
Figure 10: Comparison of running time (in second) among Rhombus, HELiKs and Cheetah for the end-to-end inference of ResNet50. The above (resp., below) image is the running time in the WAN (resp., LAN) setting.
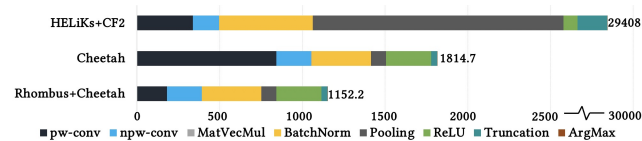
Figure 11: Comparison of the total communication (in MB) among secure two-party protocols for ResNet50 inference.

ACKNOWLEDGMENTS

REFERENCES

[1] Nitin Agrawal, Ali Shahin Shamsabadi, Matt J. Kusner, and Adrià Gascón. 2019. QUOTIENT: Two-Party Secure Neural Network Training and Prediction. In *ACM CCS 2019*, Lorenzo Cavallaro, Johannes Kinder, XiaoFeng Wang, and Jonathan Katz (Eds.). ACM Press, London, UK, 1231–1247. https://doi.org/10.1145/3319535.3339819

[2] Asra Ali, Tancrède Lepoint, Sarvar Patel, Mariana Raykova, Phillipp Schoppmann, Karn Seth, and Kevin Yeo. 2021. Communication-Computation Trade-offs in PIR, Michael Bailey and Rachel brownstadt (Eds.). USENIX Association.

[3] Sebastian Angel, Hao Chen, Kim Laine, and Srinath T. V. Setty. 2018. PIR with Compressed Queries and Amortized Query Processing. IEEE Computer Society.

[4] Gilad Asharov, Abhishek Jain, Adriana López-Alt, Eran Tromer, Vinod Vaikuntanathan, and Daniel Wichs. 2012. Multiparty Computation with Low Communication, Computation and Interaction via Threshold FHE. In *Advances in Cryptology - EUROCRYPT 2012 - 31st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cambridge, UK, April 15-19, 2012. Proceedings (Lecture Notes in Computer Science, Vol. 7237)*. Springer, 483–501.

[5] Shashank Balla and Farinaz Koushanfar. 2023. HELiKs: HE Linear Algebra Kernels for Secure Inference. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*. ACM, 2306–2320.

[6] Zvika Brakerski. 2012. Fully Homomorphic Encryption without Modulus Switching from Classical GapSVP. In *CRYPTO 2012 (LNCS, Vol. 7417)*, Reihaneh Safavi-Naini and Ran Canetti (Eds.). Springer, Heidelberg, Germany, Santa Barbara, CA, USA, 868–886. https://doi.org/10.1007/978-3-642-32009-5_50

[7] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. 2014. (Leveled) Fully Homomorphic Encryption without Bootstrapping. *ACM Trans. Comput. Theory* (2014).

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

[9] Nishanth Chandran, Divya Gupta, Aseem Rastogi, Rahul Sharma, and Shardul Tripathi. 2019. EzPC: Programmable and Efficient Secure Two-Party Computation for Machine Learning. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. 496–511.

[10] Chaochao Chen, Jun Zhou, Li Wang, Xibin Wu, Wenjing Fang, Jin Tan, Lei Wang, Alex X. Liu, Hao Wang, and Cheng Hong. 2021. When Homomorphic Encryption Marries Secret Sharing: Secure Large-Scale Sparse Logistic Regression and Applications in Risk Control, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM.

[11] Hao Chen, Wei Dai, Miran Kim, and Yongsoo Song. 2021. Efficient Homomorphic Conversion Between (Ring) LWE Ciphertexts, Kazue Sako and Nils Ole Tippenhauer (Eds.). Springer.

[12] Jung Hee Cheon, Andrey Kim, Miran Kim, and Yongsoo Song. 2017. Homomorphic encryption for arithmetic of approximate numbers. Springer.

[13] Geoffroy Couteau and Maryam Zarezadeh. 2022. Non-interactive Secure Computation of Inner-Product from LPN and LWE, Shweta Agrawal and Dongdai Lin (Eds.). Springer.

[14] Anders Dalskov, Daniel Escudero, and Marcel Keller. 2019. Secure Evaluation of Quantized Neural Networks. Cryptology ePrint Archive, Report 2019/131. https://eprint.iacr.org/2019/131.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 4171–4186.

[16] Junfeng Fan and Frederik Vercauteren. 2012. Somewhat Practical Fully Homomorphic Encryption. *IACR Cryptol. ePrint Arch.* (2012), 144.

[17] Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *41st ACM STOC*, Michael Mitzenmacher (Ed.). ACM Press, Bethesda, MD, USA, 169–178. https://doi.org/10.1145/1536414.1536440

[18] Kanav Gupta, Neha Jawalkar, Ananta Mukherjee, Nishanth Chandran, Divya Gupta, Ashish Panwar, and Rahul Sharma. 2023. SIGMA: Secure GPT Inference with Function Secret Sharing. *IACR Cryptol. ePrint Arch.* (2023), 1269.

[19] Shai Halevi and Victor Shoup. 2014. Algorithms in HElib, Juan A. Garay and Rosario Gennaro (Eds.). Springer.

[20] Shai Halevi and Victor Shoup. 2020. Design and implementation of HElib: a homomorphic encryption library. *IACR Cryptol. ePrint Arch.* (2020), 1481.

[21] Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, Guowen Xu, and Tianwei Zhang. 2022. Iron: Private Inference on Transformers. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.).

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[23] Xiaoyang Hou, Jian Liu, Jingyu Li, Yuhan Li, Wen jie Lu, Cheng Hong, and Kui Ren. 2023. CipherGPT: Secure Two-Party GPT Inference. Cryptology ePrint Archive, Paper 2023/1147. https://eprint.iacr.org/2023/1147.

[24] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs.CV]

[25] Hai Huang and Haoran Zong. 2023. Secure matrix multiplication based on fully homomorphic encryption. *J. Supercomput.* 79, 5 (2023), 5064–5085.

[26] Zhicong Huang, Wen jie Lu, Cheng Hong, and Jiansheng Ding. 2022. OpenCheetah: Lean and Fast Secure Two-Party Deep Neural Network Inference. Open-sourced libary is available at https://github.com/Alibaba-Gemini-Lab/OpenCheetah.

[27] Zhicong Huang, Wen-jie Lu, Cheng Hong, and Jiansheng Ding. 2022. Cheetah: Lean and Fast Secure Two-Party Deep Neural Network Inference, Kevin R. B. Butler and Kurt Thomas (Eds.). USENIX Association.

[28] Siam Umar Hussain, Mojan Javaheripi, Mohammad Samragh, and Farinaz Koushanfar. 2021. COINN: Crypto/ML Codesign for Oblivious Inference via Neural Networks. In *ACM CCS 2021*, Giovanni Vigna and Elaine Shi (Eds.). ACM Press, Virtual Event, Republic of Korea, 3266–3281. https://doi.org/10.1145/3460120.3484797

[29] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv:1602.07360 [cs.CV]

[30] Yuval Ishai, Joe Kilian, Kobbi Nissim, and Erez Petrank. 2003. Extending Oblivious Transfers Efficiently. In *Advances in Cryptology - CRYPTO 2003, 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003, Proceedings (Lecture Notes in Computer Science, Vol. 2729)*, Dan Boneh (Ed.). Springer, 145–161.

[31] Xiaoqian Jiang, Miran Kim, Kristin E. Lauter, and Yongsoo Song. [n. d.]. Secure Outsourced Matrix Computation and Application to Neural Networks. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018*.

[32] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha P. Chandrakasan. 2018. GAZELLE: A Low Latency Framework for Secure Neural Network Inference, William Enck and Adrienne Porter Felt (Eds.). USENIX Association.

[33] Andrey Kim, Yuriy Polyakov, and Vincent Zucca. 2021. Revisiting Homomorphic Encryption Schemes for Finite Fields, Mehdi Tibouchi and Huaxiong Wang (Eds.). Springer.

[34] Sangpyo Kim, Jongmin Kim, Michael Jaemin Kim, Wonkyung Jung, John Kim, Minsoo Rhu, and Jung Ho Ahn. 2022. BTS: an accelerator for bootstrappable fully homomorphic encryption. In *ISCA '22: The 49th Annual International Symposium on Computer Architecture, New York, New York, USA, June 18 - 22, 2022*. ACM, 711–725.

[35] Nishant Kumar, Mayank Rathee, Nishanth Chandran, Divya Gupta, Aseem Rastogi, and Rahul Sharma. [n. d.]. CrypTFlow: Secure TensorFlow Inference. In *2020 IEEE Symposium on Security and Privacy, SP 2020*.

[36] Joon-Woo Lee, HyungChul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim, and Jong-Seon No. 2022. Privacy-Preserving Machine Learning With Fully Homomorphic Encryption for Deep Neural Network. *IEEE Access* 10 (2022), 30039–30054.

[37] Jian Liu, Mika Juuti, Yao Lu, and N. Asokan. 2017. Oblivious Neural Network Predictions via MiniONN Transformations, Bhavani Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu (Eds.). ACM.

[38] Wen-jie Lu, Zhicong Huang, Zhen Gu, Jingyu Li, Jian Liu, Kui Ren, Cheng Hong, Tao Wei, and Wenguang Chen. 2023. BumbleBee: Secure Two-party Inference Framework for Large Transformers. *IACR Cryptol. ePrint Arch.* (2023), 1678. https://eprint.iacr.org/2023/1678

[39] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. 2010. On Ideal Lattices and Learning with Errors over Rings, Henri Gilbert (Ed.). Springer.

[40] Pratyush Mishra, Ryan Lehmkuhl, Akshayaram Srinivasan, Wenting Zheng, and Raluca Ada Popa. 2020. Delphi: A Cryptographic Inference Service for Neural Networks, Srdjan Capkun and Franziska Roesner (Eds.). USENIX Association.

[41] Payman Mohassel and Yupeng Zhang. 2017. SecureML: A System for Scalable Privacy-Preserving Machine Learning. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 19–38.

[42] Lucien K. L. Ng and Sherman S. M. Chow. 2021. GForce: GPU-Friendly Oblivious and Rapid Neural Network Inference. In *USENIX Security 2021*, Michael Bailey

and Rachel Greenstadt (Eds.). USENIX Association, 2147–2164.

[43] Qi Pang, Jinhao Zhu, Helen Möllering, Wenting Zheng, and Thomas Schneider. 2023. BOLT: Privacy-Preserving, Accurate and Efficient Inference for Transformers. *IACR Cryptol. ePrint Arch.* (2023), 1893. https://eprint.iacr.org/2023/1893

[44] Arpita Patra, Thomas Schneider, Ajith Suresh, and Hossein Yalame. 2021. ABY2.0: Improved Mixed-Protocol Secure Two-Party Computation. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, Michael D. Bailey and Rachel Greenstadt (Eds.). USENIX Association, 2165–2182.

[45] Deevashwer Rathee, Anwesh Bhattacharya, Divya Gupta, Rahul Sharma, and Dawn Song. 2023. Secure Floating-Point Training. In *32nd USENIX Security Symposium (USENIX Security 2023)*. USENIX Association, 6329–6346.

[46] Deevashwer Rathee, Anwesh Bhattacharya, Rahul Sharma, Divya Gupta, Nishanth Chandran, and Aseem Rastogi. 2022. SecFloat: Accurate Floating-Point meets Secure 2-Party Computation. In *2022 IEEE Symposium on Security and Privacy*. IEEE Computer Society Press, San Francisco, CA, USA, 576–595. https://doi.org/10.1109/SP46214.2022.9833697

[47] Deevashwer Rathee, Mayank Rathee, Rahul Kranti Kiran Goli, Divya Gupta, Rahul Sharma, Nishanth Chandran, and Aseem Rastogi. 2021. SiRnn: A Math Library for Secure RNN Inference. IEEE.

[48] Deevashwer Rathee, Mayank Rathee, Nishant Kumar, Nishanth Chandran, Divya Gupta, Aseem Rastogi, and Rahul Sharma. 2020. CrypTFlow2: Practical 2-Party Secure Inference, Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna (Eds.). ACM.

[49] Xuanle Ren, Zhaohui Chen, Zhen Gu, Yanheng Lu, Ruiguang Zhong, Wen-Jie Lu, Jiansong Zhang, Yichi Zhang, Hanghang Wu, Xiaofu Zheng, Heng Liu, Tingqiang Chu, Cheng Hong, Changzheng Wei, Dimin Niu, and Yuan Xie. [n. d.]. CHAM: A Customized Homomorphic Encryption Accelerator for Fast Matrix-Vector Product. In *60th ACM/IEEE Design Automation Conference, DAC 2023*.

[50] M. Sadegh Riazi, Mohammad Samragh, Hao Chen, Kim Laine, Kristin E. Lauter, and Farinaz Koushanfar. 2019. XONN: XNOR-based Oblivious Deep Neural Network Inference. In *USENIX Security 2019*, Nadia Heninger and Patrick Traynor (Eds.). USENIX Association, Santa Clara, CA, USA, 1501–1518.

[51] Nikola Samardzic, Axel Feldmann, Aleksandar Krastev, Nathan Manohar, Nicholas Genise, Srinivas Devadas, Karim Eldefrawy, Chris Peikert, and Daniel Sánchez. 2022. CraterLake: a hardware accelerator for efficient unbounded computation on encrypted data. In *ISCA '22: The 49th Annual International Symposium on Computer Architecture, New York, New York, USA, June 18 - 22, 2022*. ACM, 173–187.

[52] Phillipp Schoppmann, Adrià Gascón, Mariana Raykova, and Benny Pinkas. 2019. Make Some ROOM for the Zeros: Data Sparsity in Secure Distributed Machine Learning. In *ACM CCS 2019*, Lorenzo Cavallaro, Johannes Kinder, XiaoFeng Wang, and Jonathan Katz (Eds.). ACM Press, London, UK, 1335–1350. https://doi.org/10.1145/3319535.3339816

[53] SEAL 2021. Microsoft SEAL (release 3.7). https://github.com/Microsoft/SEAL. Microsoft Research, Redmond, WA..

[54] Liyan Shen, Ye Dong, Binxing Fang, Jinqiao Shi, Xuebin Wang, Shengli Pan, and Ruisheng Shi. 2022. ABNN2: Secure Two-Party Arbitrary-Bitwidth Quantized Neural Network Predictions. Association for Computing Machinery.

[55] Xiao Wang, Alex J. Malozemoff, and Jonathan Katz. 2016. EMP-toolkit: Efficient MultiParty computation toolkit. https://github.com/emp-toolkit.

[56] Kang Yang, Chenkai Weng, Xiao Lan, Jiang Zhang, and Xiao Wang. 2020. Ferret: Fast Extension for Correlated OT with Small Communication. In *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna (Eds.). ACM, 1607–1626.

[57] Jiawen Zhang, Jian Liu, Xinpeng Yang, Yinghao Wang, Kejia Chen, Xiaoyang Hou, Kui Ren, and Xiaohu Yang. 2024. Secure Transformer Inference Made Non-interactive. *IACR Cryptol. ePrint Arch.* (2024), 136. https://eprint.iacr.org/2024/136

[58] Qiao Zhang, Chunsheng Xin, and Hongyi Wu. 2021. GALA: Greedy ComputAtion for Linear Algebra in Privacy-Preserved Neural Networks. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society.