Enhancing Privacy Preservation with Quantum Computing for Word-Level Audio-Visual Speech Recognition

Chang Wang*, Jun Du✉*, Hang Chen*, Ruoyu Wang*, Chao-Han Huck Yang†,
Jiangjiang Zhao‡, Yuling Ren‡, Qinglong Li‡ and Chin-Hui Lee†
* University of Science and Technology of China, Hefei, Anhui, PR China
† Georgia Institute of Technology, Atlanta, GA, USA
‡ China Mobile Online Services Company Limited, China

***Abstract*—In this paper, we investigate the effectiveness of using quantum machine learning for privacy protection in audio-visual speech processing. Quantum machine learning has made significant theoretical advancements and has been shown to possess natural advantages in privacy protection over conventional techniques. Here, we first apply quantum circuits to a word-level audio-visual speech recognition task. We then propose a novel metric, an inter-class intra-class similarity ratio, for measuring the privacy-protecting capabilities of quantum circuits. Finally, we conduct an in-depth analysis of the differences in privacy protection between quantum privacy methods and traditional methods, evaluating their working principles, strengths, and limitations. Experiments results on the LRW data set show that the quantum privacy-preserving approach performs well in word-level speech recognition tasks, demonstrating excellent privacy-preserving capabilities through selective retention of features.**

I. Introduction

With the rapid development of deep learning, face recognition, automatic speech recognition and audio-visual technology have been gradually applied in daily life scenarios in recent years. The sensitivity of face image data and speech data has led to growing concerns about data privacy. In this context, it is crucial to develop effective methods for protecting audio and video privatized data to comply with new privacy protection regulations, e.g. the EU General Data Protection Regulation (GDPR) [1].

There are many traditional algorithms in the field of privacy protection, such as homomorphic encryption [2] and differential privacy [3], [4]. Homomorphic encryption is a cryptographic technique based on the computational complexity theory of mathematical puzzles. This algorithm offers the property that the result of performing an operation on plaintext and then encrypting it is equivalent to the result of performing the corresponding operation on the ciphertext after encryption. Due to this desirable feature, it can maintain high recognition accuracy while protecting data privacy. However, its application scenarios are greatly limited due to the large amount of additional computation required for the encryption process. Differential privacy methods add specific noise to the input data to prevent attackers from inferring sensitive information based on the output. However, in practical applications, it is often caught in the dilemma of trade-off between privacy-preserving capability and downstream task accuracy. To address the above challenges, [5] proposed a differential privacy method based on frequency domain transform and learnable privacy budget, which demonstrated outstanding results in face recognition tasks.

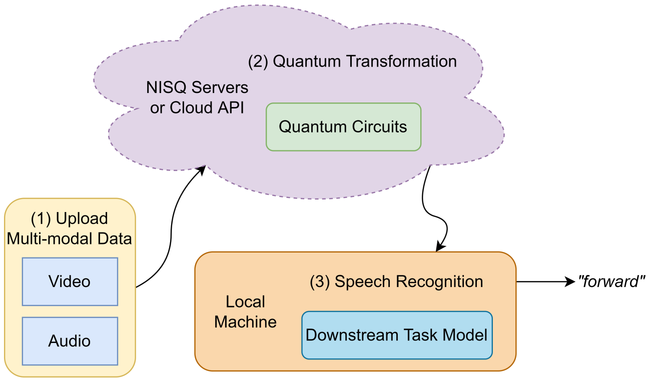

Fig. 1: Quantum privacy-preserving architecture for speech recognition task in a vertical federated learning process. Quantum circuits on the Noisy Intermediate-Scale Quantum (NISQ) server transform the input multi-modal data into quantum private data and transfer it to the speech recognition model on local machine.

In recent years, federated learning has also emerged as a common privacy-preserving approach. For example, vertical federated learning (VFL) [6] is a potential data protection strategy that decentralizes the end-to-end deep learning framework and separates feature extraction from the downstream task system. And with the recent advances in commercial quantum technologies, quantum machine learning (QML) [7] becomes an ideal building block for VFL due to its parameter encryption and data isolation advantages. The distinction between quantum machine learning and traditional machine learning lies in the utilization of quantum circuits. These circuits consist of a series of quantum gates that operate on quantum bits (qubits), which are the fundamental units of

✉ : Corresponding author

quantum information. The VFL architecture using quantum circuits as a building block are shown in Fig. 1. Based on the above, Yang et al. [8] proposed a decentralized acoustic modeling scheme and designed a quantum convolutional neural network, resulting in impressive outcomes by combining the learning paradigm of variable quantum circuit [7] and deep neural network.

The contributions of this paper are as follows:

- Demonstrates the effectiveness of quantum circuits-based methods for privacy preservation in a word-level audio-visual speech recognition task.
- Proposes the inter-class intra-class similarity ratio as a privacy protection effect metric, and use it to thoroughly analyze the differences between quantum privacy methods and traditional methods.

II. RELATED WORK

A. *Quantum Machine Learning for Signal Processing*

The current phase of quantum computing is often referred to as the noisy intermediate-scale quantum era. During this era, quantum processors consist of intermediate-scale qubits, which are not yet advanced enough to achieve fault-tolerance or reach the threshold for quantum supremacy. These processors are sensitive to external factors (noisy) and susceptible to quantum decoherence and not capable of continuous quantum error correction, which limits the potential applications of NISQ technology. However, Mitarai *et al.* [7] build a pioneering framework for building machine learning models on NISQ devices by the use of variable quantum circuit (VQC) [9]. VQC can be progressively iteratively optimized, allowing noise effects in NISQ devices to potentially be absorbed into these learned circuit parameters. Several successful machine learning applications based on VQC have been reported in the recent literature, such as transfer learning [10] and deep reinforcement learning [11]. It is worth noting that recent studies [12], [13] have shown that quantum machine learning (QML) has advantages over classical machine learning in terms of lower memory storage, secure encryption of model parameters, and good feature representation capabilities.

While still a nascent technology, quantum machine learning has already been applied in various fields. For instance, [14] successfully implemented image recognition using quantum deep convolutional networks, [15] proposed a speech recognition system using quantum backpropagation (QBP) simulated by fuzzy logic computation, and [16] successfully applied quantum transfer learning techniques to the synthetic speech detection task.

B. *Privacy Preservation*

In the realm of privacy protection, research can be broadly classified into three categories, depending on how the input data is processed: data anonymization methods, data encryption methods, and data perturbation methods. Data anonymization methods, such as k-anonymity [17] and t-closeness [18], work by desensitizing sensitive information to protect privacy. While these methods allow data to be analyzed and utilized to some extent, they may also undermine the integrity and accuracy of the data, reducing its validity. Homomorphic encryption and secure multi-party computation [19] are classical data encryption methods that effectively protect data security while achieving excellent results on downstream tasks. However, the high computational costs associated with encryption and decryption make these methods difficult to apply in practice. In contrast, data perturbation methods avoid the high costs of encryption by perturbing the original input. Differential privacy is a widely used data perturbation method. [20] employed differential privacy to perturb eigenfaces and distributed the privacy budget equally to each eigenface, which resulted in a significant loss of accuracy. Besides, [21] uses the Mixup [22] method to perturb the data, and [23] proposed a Gaussian noise perturbation method to suppress the unimportant pixels before sending them to the cloud. Recent studies [24] have started to investigate how to incorporate differential privacy into speech processing, but these efforts have not extended into the multi-modal learning setting. In summary, privacy concerns regarding audio-visual patterns deserve further investigation.

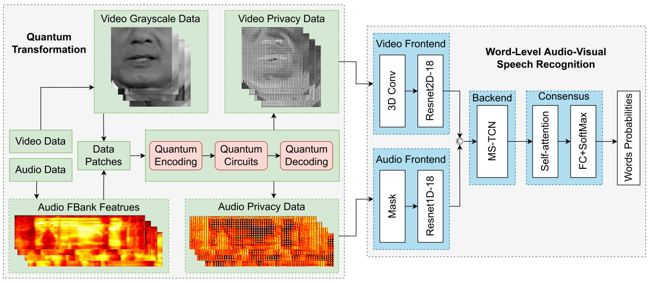

Fig. 2: Pipeline of quantum privacy-preserving method for word-level audio-visual speech recognition. © denotes the concatenate operation.

C. *Word-Level Audio-Visual Speech Recognition*

Traditional Audio-Visual Speech Recognition (AVSR) methods follow a two-step process that involves feature extraction and recognition [25]. However, recent advancements have significantly improved the performance of Visual Speech Recognition (VSR) and Automatic Speech Recognition (ASR) by combining feature extraction and recognition inside deep neural networks. In VSR, [26] proposed an end-to-end network called Visual to Phoneme. [27] proposed Spatio-Temporal Fusion Module to maintain the local spatial information and reduce the feature dimensions. [28] designs a novel deep learning architecture using hierarchical pyramidal convolution and self-attention, thereby enhancing the model's ability to discover fine-grained lip movements. In ASR, [29] demonstrated the superiority of deep representations of the network over handcrafted features such as filterbank features.

III. METHOD

A. *Audio-Visual Quantum Privacy Protection*

Figure 2 illustrates the pipeline of the audio-visual quantum privacy protection method. It comprises two modules: quantum transformation and word-level audio-visual speech recognition. In the quantum transformation module, we pre-process the audio-video data first. For audio data, we extract its FBank features; for video data, we read its grayscale image and split it into different frames for subsequent processing. Next, we chunk the processed audio and video data into patches and feed them into a quantum gate for quantum encoding into the initial quantum states. We then perform further quantum transformations using quantum circuits. Finally, the quantum state is decoded by a measurement function to output quantum privatized data.

For the word-level audio-visual speech recognition task, referring to the system in [28], [30], our system utilizes a 3D convolutional layer and an 18-layer residual network (ResNet-18) [31] as the frontend for the video modality, a mask module and a ResNet-18 as the frontend for the audio modality. The features extracted by the two frontends are concatenated and fed into a multi-scale temporal convolutional network (MS-TCN) backend, which then goes through the self-attention based consensus module to output the words probabilities.

B. Quantum Gates

In this section, we briefly introduce several basic quantum gates: R_x, R_y, R_z, and $CNOT$. Unlike a classical bit, which has two states (0 and 1), a quantum bit can be represented as $|\phi\rangle = \alpha|0\rangle + \beta|0\rangle, |\alpha|^2 + |\beta|^2 = 1$. We can also express it in the form $|\phi\rangle = \cos\theta|0\rangle + e^{-i\phi}\sin\theta|1\rangle, \theta, \phi \in R$. The density matrix of this quantum bit is:

$$\rho = |\varphi\rangle\langle\varphi| = \begin{bmatrix} \cos^2\frac{\theta}{2} & e^{-i\phi}\sin\frac{\theta}{2}\cos\frac{\theta}{2} \\ e^{i\phi}\sin\frac{\theta}{2}\cos\frac{\theta}{2} & \sin^2\frac{\theta}{2} \end{bmatrix}. \quad (1)$$

$\langle\phi|$ represents the conjugate transpose of $|\phi\rangle$.

Alternatively, we can construct the density matrix space using the four Pauli matrices as its basis, which is:

$$I = \sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_x = \sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$
$$\sigma_y = \sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_z = \sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Any matrix can be decomposed into $\rho = \sum_{i=0}^{4} a_i\sigma_i$. Considering $Tr\rho = \sum_{i=0}^{4} a_i Tr\sigma_i$, we can determine that $a_0 = 1/2$. Therefore, the density matrix can be expressed as:

$$\rho = \frac{1}{2}(I + \vec{r}\cdot\sigma) = \rho(\vec{r}),$$
$$\vec{r} = 2(a_1, a_2, a_3) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta).$$

Substituted the expression of $\vec{r}$ into the density matrix yields:

$$\rho = \frac{1}{2}\left(I + \begin{bmatrix} \cos\theta & e^{-i\phi}\sin\theta \\ e^{-i\phi}\sin\theta & -\cos\theta \end{bmatrix}\right)$$
$$= \begin{bmatrix} \cos^2\frac{\theta}{2} & e^{-i\phi}\sin\frac{\theta}{2}\cos\frac{\theta}{2} \\ e^{i\phi}\sin\frac{\theta}{2}\cos\frac{\theta}{2} & \sin^2\frac{\theta}{2} \end{bmatrix}.$$

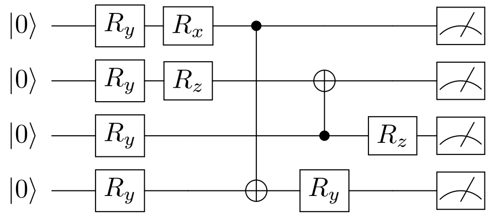

Fig. 3: An example of random quantum circuit.

This is equal to Equation 1, thus we establish the mapping relation from the quantum state to the 3D sphere, commonly known as the Bloch sphere [32].

Next, we define the quantum gates R_x, R_y, R_z to mean the representation of the quantum bit in the Bloch sphere rotating around the x, y, z axes respectively. It is then straightforward to determine their corresponding matrices as:

$$R_X(\theta) = \begin{bmatrix} \cos\frac{\theta}{2} & -i\sin\frac{\theta}{2} \\ -i\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix}, R_Y(\theta) = \begin{bmatrix} \cos\frac{\theta}{2} & -\sin\frac{\theta}{2} \\ \sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{bmatrix},$$
$$R_X(\theta) = \begin{bmatrix} e^{-i\frac{\theta}{2}} & 0 \\ 0 & e^{-i\frac{\theta}{2}} \end{bmatrix}.$$

Similarly, with reference to the Exclusive-OR gate in classical logic gates, we can express the $CNOT$ gate as a matrix:

$$CNOT = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

C. Quantum Circuits

For better privacy protection effect, the structure and parameters of quantum circuits are randomly generated. Fig. 3 gives a diagram of a random quantum circuit, where only the most basic quantum gates R_x, R_y, R_z, and $CNOT$ are applied. The classical vectors are initially encoded into a quantum state $|\phi\rangle = |0000\rangle$, and the encoded quantum states go through the quantum circuit for the following steps:

Step 1:$\Phi_1 = R_y|0\rangle R_y|0\rangle R_y|0\rangle R_y|0\rangle$.
Step 2:$\Phi_2 = R_xR_y|0\rangle R_zR_y|0\rangle R_y|0\rangle R_y|0\rangle$.
Step 3:$\Phi_3 = R_xR_y|0\rangle R_zR_y|0\rangle R_y|0\rangle CNOT\left(R_y|0\rangle\right)$.
Step 4:$\Phi_4 = R_xR_y|0\rangle CNOT\left(R_zR_y|0\rangle\right) R_y|0\rangle R_y \text{ CNOT}\left(R_y|0\rangle\right)$.
Step 5:$\Phi_5 = R_xR_y|0\rangle CNOT\left(R_zR_y|0\rangle\right) R_zR_y|0\rangle R_y CNOT\left(R_y|0\rangle\right)$.

Furthermore, due to the potential for many unexpected noisy signals from $CNOT$ gates, particularly with our current non-error-corrected quantum devices and the connectivity of physical qubits, we limit the number of qubits in our random quantum circuit to ensure it falls within the noise tolerance capabilities of VQC.

IV. Privacy Protection Capability Measurement

A. Inter-class Intra-class Similarity Ratio

To quantitatively analyze the privacy protection capability of different methods, we introduce a new metric called the inter-class intra-class similarity ratio (IISR).

To process video data, we group frames from the same video together as one class, while frames from different videos are treated as separate classes. Audio data is handled in a similar fashion. Suppose there are n video segments in total, and each video segment has m frames, the j-th frame image of the i-th video segment is $\mathrm{fig}_{i,j}$.

First, we utilize SSIM as the measure of the similarity between data. From there, we can calculate the intra-class similarity by employing the following formula:

$$\mathrm{ssim}_{\text{intra}} = \frac{1}{n * m * (m-1)} \sum_{i=1}^{n} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} \mathrm{ssim}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{i,k}\right).$$

Similarly, the inter-class similarity function is:

$$\mathrm{ssim}_{\text{inter}} = \frac{1}{n * (n-1) * m^2} \sum_{\substack{i,l=1 \\ i \neq l}}^{n} \sum_{j,k=1}^{m} \mathrm{ssim}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{l,k}\right).$$

To calculate the IISR of SSIM, we can express it as the ratio of inter-class similarity to intra-class similarity:

$$\mathrm{IISR}_{\text{ssim}} = \frac{(m-1) \sum_{\substack{i,l=1 \\ i \neq l}}^{n} \sum_{j,k=1}^{m} \mathrm{ssim}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{l,k}\right)}{m * (n-1) \sum_{i=1}^{n} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} \mathrm{ssim}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{i,k}\right)}.$$

Similarly, the formulas for the IISR of RMSE and PSNR are:

$$\mathrm{IISR}_{\text{rmse}} = \frac{(m-1) \sum_{\substack{i,l=1 \\ i \neq l}}^{n} \sum_{j,k=1}^{m} 1/\mathrm{rmse}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{l,k}\right)}{m * (n-1) \sum_{i=1}^{n} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} 1/\mathrm{rmse}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{i,k}\right)}.$$

$$\mathrm{IISR}_{\text{psnr}} = \frac{(m-1) \sum_{\substack{i,l=1 \\ i \neq l}}^{n} \sum_{j,k=1}^{m} \mathrm{psnr}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{l,k}\right)}{m * (n-1) \sum_{i=1}^{n} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} \mathrm{psnr}\left(\mathrm{fig}_{i,j}, \mathrm{fig}_{i,k}\right)}.$$

Inter-class and intra-class similarities respectively denote the proportion of shared information contained within inter-class and intra-class data. Greater similarity indicates more compact data. The IISR represents the proportion of common information contained within inter-class and intra-class data. A smaller ratio implies that the intra-class data is more compact than the inter-class data, making it easier to differentiate whether the two data belong to the same class, which in turn means more information is preserved.

B. Privacy Attack

To assess the effectiveness of privacy protection methods, privacy attacks [24] are often employed. It can be broadly classified into white-box attacks and black-box attacks.

1) White-box attack: White-box attacker has access to all of our operations. With this knowledge, in order to attack the quantum privacy protection method, they can take all parameters in the quantum circuits and inverse them to get inverse quantum circuits, and then feed the quantum privatized data into the inverse quantum circuits to get the output.

2) Black-box attack: A black-box attack assumes that the attacker does not possess any knowledge regarding the model's internal structure and parameters. Nevertheless, attackers can collect large amounts of audio and video data and input the data into the model to obtain processed data. Subsequently, they can train a decoder to restore the processed results to the original input. Finally, attackers can leverage the trained decoder to recover users' private data. For our subsequent experiments, we will use UNet [33] as our decoder to reconstruct the original data from the processed data.

V. Experiments

A. Experimental Setup

To evaluate the impact of two privacy-preserving methods on downstream tasks, we selected a moderately challenging yet feasible task: word-level speech recognition. For the dataset, we use the Lip Reading in the Wild (LRW) [34], which consists of about 500,000 audio-visual speech segments extracted from BBC TV broadcasts. It encompasses 500 target words and numerous speakers, and provides a word-level label for each audio-visual speech segment.

To account for practical application factors, we introduce noise to the original speech. A total of 115 noise types including 100 noise types from [35] and 15 homemade noise types corrupted the corpus in the training set at 5 SNR levels (i.e., 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB). The validation and test sets were corrupted by three unseen noise types at the aforementioned SNR levels, i.e., Speech Babble, Buccaneer1, and Destroyer Engine. All unseen noises were collected from the NOISEX92 corpus [36]. Following the addition of noise, we extracted its 40-dim Fbank features from the noisy speech as the input. Regarding the video modality, we utilize grayscale images directly as input.

We employed the method proposed by [5] as the baseline for privacy protection. In the frequency domain transform part, we use some functions in TorchJPEG [37] to transform the input data into BDCT coefficients and then remove its direct part. And for the initial value of the learnable budget allocation parameter, we set it to 0 so that the privacy budget of each location is equal in the initial stage.

For the quantum privacy-preserving method, we use Pennylane to build a 4-bits random quantum circuits and divide the input data into 2×2 patches into the quantum circuits to generate quantum privatized data.

B. Performance of Word-level Audio-visual Speech Recognition

We conducted a comparative analysis of the classification accuracy for speech recognition task using original data, baseline

TABLE I: Speech recognition accuracy(Acc) in different data types and modalities.

Acc	Original	Baseline	**Quantum**
ASR	69.00	65.92	**68.12**
VSR	86.43	84.23	**85.42**
AVSR	89.23	87.13	**88.26**

privatized data, and quantum privatized data, respectively. The results are shown in Table I.

Based on the table, it is apparent that the classification accuracy of the original data is the highest, irrespective of the modality. This implies that both privacy-preserving methods lead to some information loss. Nonetheless, we observed that the training results of the quantum privatized data are superior to the baseline approach, suggesting that the quantum privacy-preserving method loses less information for the downstream task required.

C. privatized data Analysis

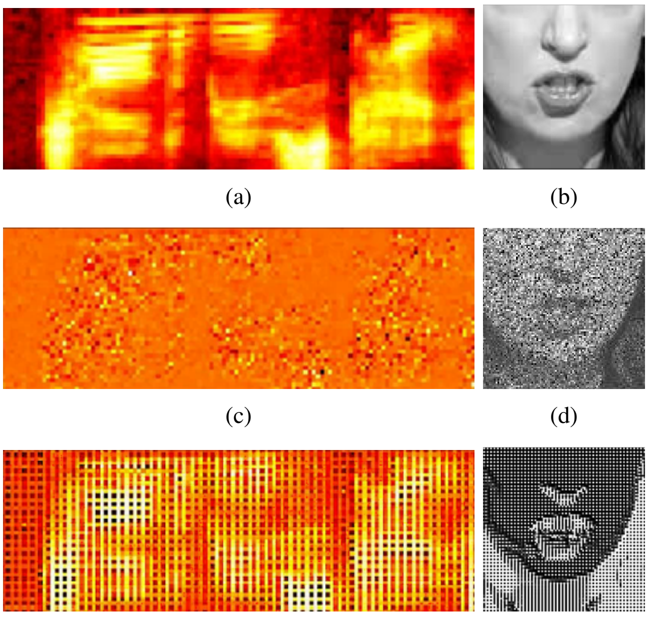
(a) (b) (c) (d) (e) (f)

Fig. 4: Visualization of privatized data. (a) Spectrogram of original audio data; (b) Grayscale plot of original video data; (c) Spectrogram of audio baseline privatized data; (d) Grayscale plot of video baseline privatized data; (e) Spectrogram of audio quantum privatized data; (f) Grayscale plot of video quantum privatized data.

As the first step of our analysis, we performed visualization operations on the dataset. Specifically, We randomly select an audio-video segments from the original dataset, generated the grayscale map of the first frame of the video and the audio spectrogram, and repeated the same visualization procedure for the baseline privatized data and quantum privatized data corresponding to this segments. The resulting visualizations are shown in Fig. 4.

TABLE II: privatized data analysis.

Modality	Data_type	IISR_{rmse}	IISR_{psnr}	IISR_{ssim}
Video	Original	0.314	0.611	0.491
	Baseline	0.532	0.741	0.585
	Quantum	**0.539**	**0.772**	**0.680**
Audio	Original	0.611	0.824	0.407
	Baseline	0.855	0.925	0.675
	Quantum	**0.926**	**0.973**	**0.776**
Video_lip	Original	0.457	0.738	0.533
	Baseline	**0.693**	**0.856**	**0.663**
	Quantum	0.641	0.818	0.635

TABLE III: Data analysis of white-box attack in video modality.

Data_type	IISR_{rmse}	IISR_{psnr}	IISR_{ssim}
Original	0.314	0.611	0.491
Quantum	0.539	0.772	0.680
White-box Attack	0.536	0.761	0.662

As per our analysis of the visualizations, we observed that the quantum privatized data the quantum privatized data better preserves the lip area compared to the baseline data while worse in other area. Similarly, we noted that quantum privatized data captures high-frequency information better than the baseline privatized data from the audio spectrograms.

For further analysis, we employ the IISR metric. We randomly select 10 words and 30 audio-video segments for each word, and calculate IISR for the audio and video data separately. Moreover, given the significant reliance of the VSR task on the lip area in the video, we extracte the lip area of the video segments and compute its IISR. The results are shown in Table II.

From the first two rows of the table, we can see that for both the video and audio data, the IISR values for the original, baseline, and quantum privatized data follow the trend of $Original_{\text{IISR}} < Baseline_{\text{IISR}} < Quantum_{\text{IISR}}$, regardless of the IISR_{rmse}, IISR_{psnr} or IISR_{ssim}. This suggests that the quantum privatized data retains relatively less information compared to the baseline method, thereby providing stronger privacy protection. By further analyzing the IISR of lip area data, we can find that there is $Original_{\text{IISR}} < Quantum_{\text{IISR}} < Baseline_{\text{IISR}}$, which is different from previous results and indicating that quantum privatized data retain more information in the lip region, thus performing better in downstream tasks.

D. White-box Attacking Experiments

The white-box attack results is visualized in Fig. 5. Additionally, we calculated the IISR for the white-box attack results, with the results shown in Table III.

As we can see from the figure or the table, the impact of white-box attacks is very limited. The quantum privacy protection method exhibited strong resistance to white-box

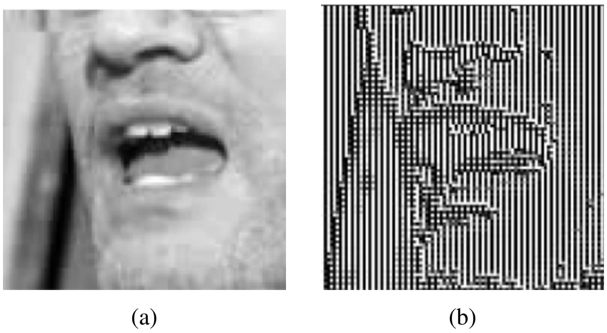

Fig. 5: Visualization of white-box attack. (a) Grayscale plot of original video data; (b) Grayscale plot of white-box attack results of video quantum privatized data.

attacks, which is indicative of the irreversibility of quantum circuits.

E. Black-box Attacking Experiments

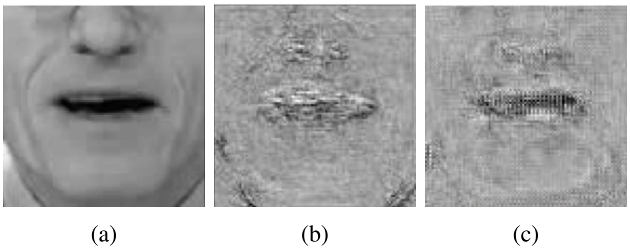

Fig. 6: Visualization of untrained black-box attack. (a) Grayscale plot of original video data; (b) Grayscale plot of untrained black-box attack results of video baseline privatized data; (c) Grayscale plot of untrained black-box attack results of video quantum privatized data.

In this section, we discuss the privacy-preserving reliability of the two approaches under black-box attacks.

Since the quantum privacy method does not require any trainable parameters and the number of parameters required for the baseline method is small, we examined the reconstruction results of the untrained attack network after initialization. The results are visualized in Fig. 6. From the figure, it is evident that compared to the baseline, the quantum privatized data reconstruction results are better restored near the lips, but worse restored in other regions.

The visualization results of the audio and video data reconstructed by the trained black-box attack network are also shown in Fig. 7. The video visualizations reveal that while both the quantum privacy protection method and the baseline method are effective at resisting black-box attacks, the restoration results for the video quantum privatized data are superior to those of the video baseline privatized data in the lip region. Similarly, the audio quantum privatized data exhibits better restoration capability for high-frequency information compared to the audio baseline privatized data.

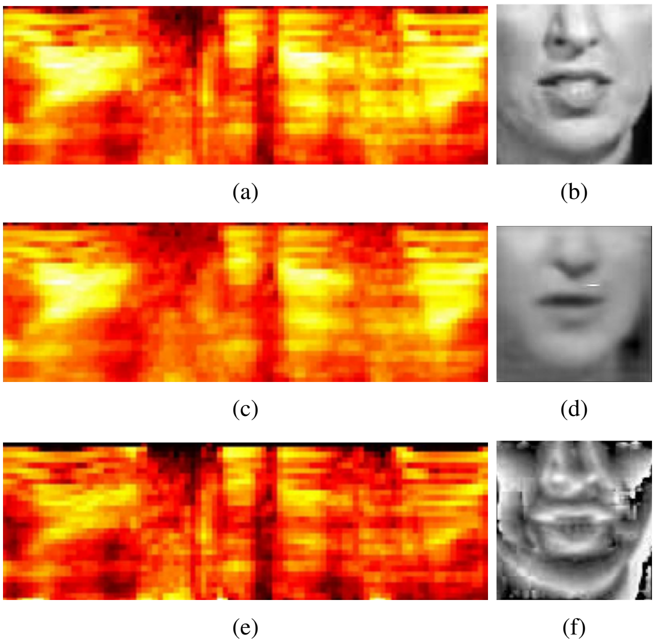

Fig. 7: Visualization of black-box attack. (a) Spectrogram of original audio data; (b) Grayscale plot of original video data; (c) Spectrogram of black-box attack results of audio baseline privatized data; (d) Grayscale plot of black-box attack results of video baseline privatized data; (e) Spectrogram of black-box attack results of audio quantum privatized data; (f) Grayscale plot of black-box attack results of video quantum privatized data.

TABLE IV: Data analysis of black-box attack.

Modality	Data_type	IISR$_{rmse}$	IISR$_{psnr}$	IISR$_{ssim}$
Video	Original	0.314	0.611	0.491
	Baseline	0.489	0.838	0.773
	Quantum	**0.632**	**0.867**	**0.892**
Audio	Original	0.611	0.824	0.407
	Baseline	0.798	0.956	0.882
	Quantum	**0.914**	**0.984**	**0.947**
Video_lip	Original	0.457	0.738	0.533
	Baseline	**0.589**	**0.864**	**0.894**
	Quantum	0.562	0.839	0.848

We compute the IISR values of the reconstructed audio and video data, with the results depicted in Table IV. From the table, we can see that for both audio and video modalities, the reconstruction results of quantum privatized data are inferior to those of the baseline privatized data, indicating that the quantum protection approach is more resistant to black-box attack than the baseline. Furthermore, we analyze the lip region data separately and find that the reconstruction results of quantum privatized data are better than those of the baseline privatized data. This observation provides further evidence to support the superior performance of the quantum privatized data for word-level speech recognition tasks.

VI. Conclusions

This paper successfully demonstrates the effectiveness of quantum circuits-based privacy-preserving methods in audio and video modalities. Through visualizations and the IISR metric, we are able to analyze the privacy-preserving ability of the two methods in terms of privatized data, resistance to white-box privacy attacks, and black-box privacy attacks. Furthermore, by analyzing the data in the lip region, we can find that the quantum privacy-preserving methods can maintain the features required for downstream tasks while discarding other information, thereby enabling excellent privacy-preserving capabilities while performing well in downstream tasks, which is inseparable from the advantage of quantum circuits in feature representation capability.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants No. 62171427.

References

[1] P. Voigt and A. V. dem Bussche, “The eu general data pro tection regulation (gdpr),” *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[2] C. Gentry and S. Halevi, “Implementing gentry’s fully-homomorphic encryption scheme,” in *International Conference on the Theory and Application of Cryptographic Techniques*, 2011.

[3] C. Dwork, “Differential privacy: A survey of results,” in *Theory and Applications of Models of Computation*, 2008.

[4] C.-H. H. Yang, S. M. Siniscalchi, and C.-H. Lee, “Pate-aae: Incorporating adversarial autoencoder into private aggregation of teacher ensembles for spoken command classification,” *Proc. of Interspeech*, 2021.

[5] J.-B. Ji, H. Wang, Y. Huang, *et al.*, “Privacy-preserving face recognition with learnable privacy budgets in frequency domain,” in *European Conference on Computer Vision*, 2022.

[6] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *arXiv: Artificial Intelligence*, 2019.

[7] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, “Quantum circuit learning,” *Physical Review A*, 2018.

[8] C.-H. H. Yang, J. Qi, S. Y.-C. Chen, *et al.*, “Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition,” *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6523–6527, 2020.

[9] M. Benedetti, E. Lloyd, S. H. Sack, and M. Fiorentini, “Parameterized quantum circuits as machine learning models,” *Quantum Science and Technology*, vol. 4, 2019.

[10] J. Qi and J. Tejedor, “Classical-to-quantum transfer learning for spoken command recognition based on quantum neural networks,” *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8627–8631, 2021.

[11] S. Y.-C. Chen, C.-H. H. Yang, J. Qi, P.-Y. Chen, X. Ma, and H.-S. Goan, “Variational quantum circuits for deep reinforcement learning,” *IEEE Access*, vol. 8, pp. 141 007–141 024, 2019.

[12] V. Havlíek, A. D. Córcoles, K. Temme, *et al.*, “Supervised learning with quantum-enhanced feature spaces,” *Nature*, vol. 567, pp. 209–212, 2018.

[13] C.-H. H. Yang, B. Li, Y. Zhang, *et al.*, “A quantum kernel learning approach to acoustic modeling for spoken command recognition,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[14] Y. Li, R.-g. Zhou, R. G. Xu, J. Luo, and W. Hu, “A quantum deep convolutional neural network for image recognition,” *Quantum Science & Technology*, vol. 5, 2020.

[15] F. Li, S. Zhao, and B.-y. Zheng, “Quantum neural network in speech recognition,” *6th International Conference on Signal Processing, 2002.*, vol. 2, 1267–1270 vol.2, 2002.

[16] R. Wang, J. Du, and T. Gao, “Quantum transfer learning using the large-scale unsupervised pre-trained model wavlm-large for synthetic speech detection,” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[17] L. Sweeney, “K-anonymity: A model for protecting privacy,” *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, vol. 10, pp. 557–570, 2002.

[18] N. Li, T. Li, and S. Venkatasubramanian, “T-closeness: Privacy beyond k-anonymity and l-diversity,” *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115, 2007.

[19] F. Boemer, R. Cammarota, D. Demmler, T. Schneider, and H. Yalame, “Mp2ml: A mixed-protocol machine learning framework for private inference,” *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 2020.

[20] P. C. M. Arachchige, P. Bertók, I. Khalil, D. Liu, and S. A. Çamtepe, “Privacy preserving face recognition utilizing differential privacy,” *Comput. Secur.*, vol. 97, p. 101 951, 2020.

[21] G. McGraw and A. K. Ghosh, “Developing expertise in software security an outsider s perspective,” 1996.

[22] H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz, “Mixup: Beyond empirical risk minimization,” *ArXiv*, vol. abs/1710.09412, 2017.

[23] F. Mireshghallah, M. Taram, A. Jalali, A. T. Elthakeb, D. M. Tullsen, and H. Esmaeilzadeh, “Not all features are equal: Discovering essential features for preserving

prediction privacy," *Proceedings of the Web Conference 2021*, 2021.

[24] C.-H. H. Yang, I.-F. Chen, A. Stolcke, S. M. Siniscalchi, and C.-H. Lee, "An experimental study on private aggregation of teacher ensemble learning for end-to-end speech recognition," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 1074–1080.

[25] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multim.*, vol. 2, pp. 141–151, 2000.

[26] B. Shillingford, Y. Assael, M. W. Hoffman, *et al.*, "Large-scale visual speech recognition," *ArXiv*, vol. abs/1807.05162, 2018.

[27] X. Zhang, F. Cheng, and S. Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 713–722, 2019.

[28] H. Chen, J. Du, Y. Hu, L.-R. Dai, B.-C. Yin, and C.-H. Lee, "Automatic Lip-Reading with Hierarchical Pyramidal Convolution and Self-Attention for Image Sequences with No Word Boundaries," in *Proc. Interspeech 2021*, 2021, pp. 3001–3005.

[29] T. Parcollet, M. Morchid, and G. Linarès, "E2e-sincnet: Toward fully end-to-end speech recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7714–7718, 2020.

[30] B. Martínez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6319–6323, 2020.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[32] F. Bloch, N. INTRODUCTIO, R. Zacharias, *et al.*, "Nuclear induction," 2011.

[33] J. Qian, R. Li, X. Yang, *et al.*, "Hasa: Hybrid architecture search with aggregation strategy for echinococcosis classification and ovary segmentation in ultrasound images," *ArXiv*, vol. abs/2204.06697, 2022.

[34] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[35] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 2067–2079, 2010.

[36] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[37] M. Ehrlich, S.-N. Lim, L. S. Davis, and A. Shrivastava, "Quantization guided jpeg artifact correction," in *European Conference on Computer Vision*, 2020.