Privacy-Preserving Multimodal Sentiment Analysis

Honghui Xu, *Member, IEEE*, Wei Li, *Member, IEEE*, Daniel Takabi, *Member, IEEE*, Daehee Seo, *Member, IEEE*, and Zhipeng Cai, *Fellow, IEEE*

***Abstract*—Multimodal sentiment analysis plays a critical role in numerous IoT-driven applications, such as personalized smart assistants, healthcare monitoring systems, and intelligent transportation networks, where accurate interpretation of user emotions is vital for enhancing service quality. However, a severe threat of privacy leakage in the multimodal sentiment analysis has been overlooked by previous works. To fill this gap, we propose a differentially private correlated representation learning (DPCRL) model to achieve privacy-preserving multimodal sentiment analysis by combining a correlated representation learning scheme with a differential privacy protection scheme. Our correlated representation learning scheme aims to achieve heterogeneous multimodal data transformation to meet the requirements of privacy-preserving multimodal sentiment analysis by learning the correlated and uncorrelated representations, where especially, a predetermined correlation factor is employed to flexibly adjust the expected correlation among the correlated representations. The differential privacy protection scheme is used to obtain the disturbed correlated and uncorrelated representations by adding Laplace noise for ϵ-differential privacy. In particular, the correlation factor can help alleviate the side-effect of the added Laplace noise on the sentiment prediction performance. Finally, via conducting a series of real-data experiments, we validate that our proposed DPCRL model is superior to the state-of-the-art for privacy-preserving multimodal sentiment analysis.**

***Index Terms*—Differential privacy, multimodal systems, representation learning, sentiment analysis.**

I. Introduction

WITH the proliferation of smart infrastructures in IoT applications, multimodal sentiment analysis has become increasingly important for enhancing user interactions in various scenarios, such as smart homes [1], healthcare systems [2], and intelligent transportation [3]. Driven by advancements in deep learning, learning-based prediction has emerged as a promising and effective approach for realizing multimodal sentiment analysis through the integration of multimodal data representations extracted from raw multimedia inputs [4], [5], [6]. However, in IoT contexts where devices continuously generate sensitive user data, these extracted representations can be exploited by malicious attackers to infer private information (e.g., user identity, behavioral patterns, and location), leading to significant privacy risks and potential economic losses [7], [8], [9], [10], shown in Fig. 1. This underscores the critical need for privacy-preserving mechanisms specifically tailored for multimodal sentiment analysis in IoT scenarios, where ensuring data security while maintaining system efficiency is paramount. To address this, our work focuses on designing robust privacy-preserving models that are applicable to real-world IoT deployments, offering a secure foundation for multimodal sentiment analysis.

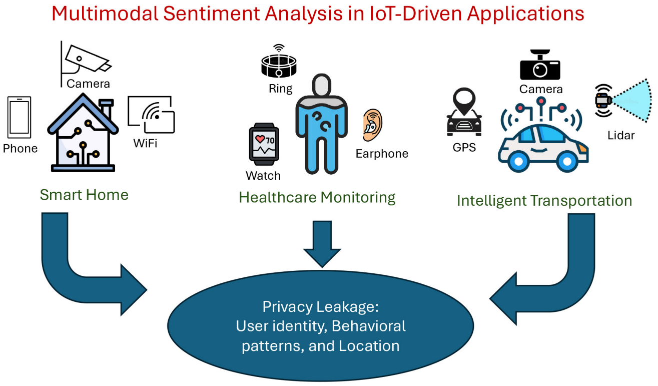
Multimodal Sentiment Analysis in IoT-Driven Applications
Camera
Phone
WiFi
Smart Home
Ring
Watch
Earphone
Healthcare Monitoring
Camera
GPS
Lidar
Intelligent Transportation
Privacy Leakage:
User identity, Behavioral patterns, and Location

Fig. 1. Privacy leakage in multimodal sentiment analysis in IOT-driven smart infrastructures.

In order to prevent privacy leakage from learning-based multimodal sentiment analysis methods, a number of privacy-preserving learning algorithms have been proposed [11], [12], [13]. One vein of research is based on adversarial training to generate adversarial samples that is used as the data disturbed by noise to defend inference attacks not only on unimodal data [14], [15] but also on multimodal data [16], [17], [18]. Although these adversarial training-based models are widely applied to privacy-preserving learning schemes, they fail to provide any performance guarantee of data privacy protection. Differential privacy-based models [19], [20] have been developed to guarantee data privacy protection by disturbing the data via the addition of Laplace noise based on differential privacy mechanisms [21], [22], [23], [24], [25]. However, it is worth mentioning that the data correlation can be treated as side-channel information, thus reducing the effectiveness of differential privacy protection. As a result,

Received 21 October 2024; revised 7 December 2024; accepted 6 January 2025. Date of publication 9 January 2025; date of current version 23 May 2025. This work was supported in part by the National Science Foundation of U.S. under Grant 2416872, Grant 2315596, Grant 2244219, Grant 2146497, Grant 2343619, Grant 2413856, and Grant 2413654. *(Corresponding author: Wei Li.)*

Honghui Xu is with the Department of Information Technology, Kennesaw State University, Kennesaw, GA 30144 USA (e-mail: hxu10@kennesaw.edu).

Wei Li and Zhipeng Cai are with the Department of Computer Science, Georgia State University, Atlanta, GA 30302 USA (e-mail: wli28@gsu.edu; zcai@gsu.edu).

Daniel Takabi is with the School of Cybersecurity, Old Dominion University, Norfolk, VA 23529 USA (e-mail: takabi@odu.edu).

Daehee Seo is with the College of Intelligence Information Engineering, Sangmyung University, Seoul 03016, Republic of Korea (email: daehseo@smu.ac.kr).

Digital Object Identifier 10.1109/JIOT.2025.3527864

for correlated data, the additional Laplace noise used in differential privacy mechanisms should be enlarged with the increase of data correlation to maintain the same differential privacy protection degree, inevitably sacrificing the learning performance (*e.g.,* accuracy) [26], [27], [28]. Furthermore, to mitigate the impact of data correlation on performance loss, the existing differentially private transform-based approaches transform the correlated homogeneous data into the corresponding uncorrelated data domain and then implement differential privacy mechanisms to achieve data privacy guarantee [29], [30], [31], [32], [33]. Nevertheless, these existing transform-based approaches can only perform the transformation on homogeneous data with intracorrelation (that means data correlation within a data instance, such as temporal correlation in a video and location correlation in a trajectory) but are not applicable to heterogeneous data with intercorrelation [34], [35], [36] (that means correlation among different data instances, such as data correlation between two texts and data correlation between a video and an audio). This is because the transformation schemes in the previous works, including discrete Fourier transform (DFT), wavelet transform (WT), and principle component analysis (PCA), can only process the correlated homogeneous data to generate uncorrelated representations. Therefore, it is still a challenging task to generate privacy-preserving representations of the correlated heterogeneous multimodal data while maintaining the performance of multimodal sentiment analysis.

Motivated by the above analysis, in this article, we devise a novel model, named differentially private correlated representation learning (DPCRL), to generate privacy-preserving multimodal representations for multimodal sentiment analysis by integrating a correlated representation learning scheme and a differential privacy protection scheme. The correlated representation learning scheme is designed as a heterogeneous multimodal data transformation strategy to learn the correlated and uncorrelated multimodal representations, in which a correlated factor can be predetermined to flexibly adjust the expected correlation among the correlated multimodal representations. The differential privacy protection scheme is further applied to generating the disturbed correlated and uncorrelated representations by adding Laplace noise for satisfying ϵ-differential privacy. More specifically, a proper correlation factor can be set in our DPCRL model to extract the correlated representations with a relatively lower correlation, thus mitigating the side-effect of the additional Laplace noise on sentiment prediction performance. Finally, we evaluate the effectiveness of our DPCRL model on real-world datasets by conducting comprehensive experiments. Our multifold contributions are addressed as follows.

1) To the best of our knowledge, this is the first work to design privacy-preserving multimodal sentiment analysis model.
2) Our proposed DPCRL model seamlessly combines a correlated representation learning scheme with a differential privacy protection scheme, aiming at simultaneously ensuring ϵ-differential privacy and retaining the performance of multimodal sentiment analysis.
3) In our correlated representation learning scheme, the heterogeneous multimodal data transformation can be accomplished by learning the correlated and uncorrelated multimodal representations from multimodal data for sentiment prediction, and the expected correlation of correlated representations can be flexibly set via a correlation factor.
4) Comprehensive experiments are well conducted to validate the advantages of our DPCRL model over the state-of-the-art for privacy-preserving multimodal sentiment analysis.

The remainder of this article is organized as follows. The related works are briefly summarized in Section II. We elaborate the details of our model in Section III, and then conduct real-data experiments and analyze all the results in Section IV. Finally, we end up with a conclusion in Section V.

II. Related Works

In this section, we summarize the related works on multimodal sentiment analysis and review the current mainstream privacy-preserving learning approaches.

A. Multimodal Sentiment Analysis

The current landscape of multimodal sentiment analysis research reflects a growing interest and significant progress in the field [37]. Researchers have increasingly recognized the value of leveraging multiple modalities [38], [39], [40], [41], [42], [43], such as text, audio, and video, to capture rich and nuanced expressions of sentiment [44], [45]. A variety of approaches have been explored, ranging from traditional machine learning techniques [46], [47] to deep learning architectures [48], [49], each with its strengths and limitations. Recent studies have focused on developing more sophisticated multimodal fusion methods [50] to effectively integrate information from diverse modalities [51]. Despite these advancements, challenges, such as multimodal alignment [52] and data heterogeneity [53] persist, motivating ongoing research into novel methodologies and solutions. These approaches can be categorized into modality interaction-based, modality transformation-based, and modality similarity-based methods. *Interaction-based methods* explore dynamic interplays between modalities to utilize complementary information [52], while *transformation-based approaches* [54], [55] transform modalities into a common feature space for unified analysis. Additionally, *similarity-based schemes* [56] focus on modality correlations to enhance sentiment analysis. These foundational methodologies inform our work, where we introduce a correlation factor within a differential privacy framework, a novel integration that strategically manipulates modality correlations to balance data utility and privacy in MSA, filling a specific gap not directly addressed by existing studies.

B. Privacy-Preserving Learning Approaches

Currently, adversarial training-based models, differential privacy-based approaches, and differentially private transform-based methods are the mainly popular techniques used in machine learning for data privacy protection.

1) *Adversarial training-based models* are exploited to generate adversarial samples that are taken as the data

disturbed by noise to defend learning-based inference attacks not only for unimodal data [14], [15], [57] but also for multimodal data [17], [18]. Although the adversarial training is relatively attractive to be employed in privacy-preserving learning schemes owing to its convenience and efficiency, it cannot ensure a privacy protection guarantee.

2) *Differential privacy-based approaches* are proposed to provide a theoretical guarantee of data privacy protection by adding Laplace noise based on differential privacy mechanisms [22], [58], [59], [60], [61], [62]. In particular, for the correlated data, the added Laplace noise should be increased with the growth of data correlation so as to ensure the theoretical guarantee of data privacy protection [28], which however, sacrifices the performance (*e.g.,* accuracy) of learning models.
3) *Differentially private transform-based methods* transform the correlated data into the corresponding uncorrelated data domain and then apply differential privacy mechanisms to preserve data privacy [29], [32], where the side-effect of the larger Laplace noise on learning performance can be eliminated due to the disappearance of data correlation after data transformation.

Unfortunately, these existing transform-based methods can only be used to transform the homogeneous data with intra-correlation into independent (uncorrelated) data domain but cannot be applied to the heterogeneous multimodal data with intercorrelation.

In this article, a novel DPCRL model is proposed to ensure differential privacy while maintaining the performance of multimodal sentiment analysis. In DPCRL, the heterogeneous multimodal data transformation can be achieved by learning the correlated and uncorrelated multimodal representations, where especially, a predetermined correlation factor can be used to adjust the expected correlation of the correlated representations. More importantly, a proper correlation factor can help mitigate the side-effect of the added Laplace noise on sentiment prediction performance.

III. Methodology

In this section, we elaborate on the details of our proposed DPCRL model. As shown in Fig. 2, the DPCRL model is made up of five components, including a feature extraction module, an encoding module, a decoding module, a differential privacy protection module, and a privacy-preserving sentiment prediction module. First, a feature extraction scheme is designed to extract features from video, audio, and language modalities. Second, in the encoding module, we use the correlated and uncorrelated multimodal representation encoders to learn the correlated and uncorrelated multimodal representations from the extracted features, where a correlation factor is used in the correlated multimodal representation encoders to obtain the correlated multimodal representations. Third, the decoding module is devised to reconstruct the extracted features by decoding the correlated and uncorrelated representations in each modality, which helps the encoding module avoid encoding the unrepresentative vector in each modality. This autoencoding architecture of the correlated representation learning actually works as a heterogeneous multimodal data transform scheme in DPCRL. Fourth, a differential privacy protection scheme is leveraged to obtain privacy-preserving representations by adding Laplace noise to the correlated and uncorrelated representations learned from the previous autoencoding architecture. Finally, these perturbed representations are put into the privacy-preserving sentiment prediction module to accomplish the privacy-preserving multimodal sentiment analysis task.

For real-world implementation, the first four components in DPCRL should be deployed on the users' device side, and the final one should be implemented on the server side. When running DPCRL, the first four components are executed on the users' device side to generate the privacy-preserving representations, which will be transmitted to the server side for the final prediction using the final component. DPCRL can help users avoid privacy leakage caused by attackers who can leverage the eavesdropped representations during transmission to infer the raw users' sensitive data via some effective deep learning attack models, such as the membership inference attack and the inversion attack. In the following, we introduce these five modules in DPCRL one by one.

A. Feature Extraction

Each video is segmented into utterances, each of which is a unit of speech bounded by breaths or pauses [63]. An utterance comprises a sequence of visual modality data denoted as $\mathbf{U}_v \in \mathbb{R}^{T_v \times d_v}$, a sequence of acoustic modality data denoted as $\mathbf{U}_a \in \mathbb{R}^{T_a \times d_a}$, and a sequence of language modality data denoted $\mathbf{U}_l \in \mathbb{R}^{T_l \times d_l}$, where T_m $(m \in \{v, a, l\})$ represents the length of an utterance, and d_m represents the number of dimensions of the modality data. For feature extraction, the stacked bidirectional long short-term memory scheme (sLSTM) [64] is exploited to map $\mathbf{U}_m \in \mathbb{R}^{T_m \times d_m}$ into a feature vector $\mathbf{f}_m \in \mathbb{R}^{d_h}$ $(m \in \{v, a, l\})$ with d_h being the size of hidden states set in the sLTSM model

$$\mathbf{f}_m = \text{sLSTM}(\mathbf{U}_m; \theta_m^{\text{slstm}}) \tag{1}$$

where θ_m^{lstm} represents the parameters of sLSTM.

B. Encoding

In the encoding process, the visual/acoustic/language modality data is processed by taking into account the following three requirements.

1) For each feature vector $\mathbf{f}_m$ $(m \in \{v, a, l\})$, its correlated and uncorrelated representations should capture two distinctive aspects of the same modality data.
2) Any two of the uncorrelated representations of $\mathbf{f}_v$, $\mathbf{f}_a$, and $\mathbf{f}_l$ should be distinctive without redundancy.
3) The correlation between any two of the correlated representations of $\mathbf{f}_v$, $\mathbf{f}_a$, and $\mathbf{f}_l$ should be close to the correlation factor c as much as possible.

First, as shown by domain separation networks [65], each feature vector $\mathbf{f}_m$ can be projected to two distinct types of representations. Thus, given $\mathbf{f}_m$, we use the correlated

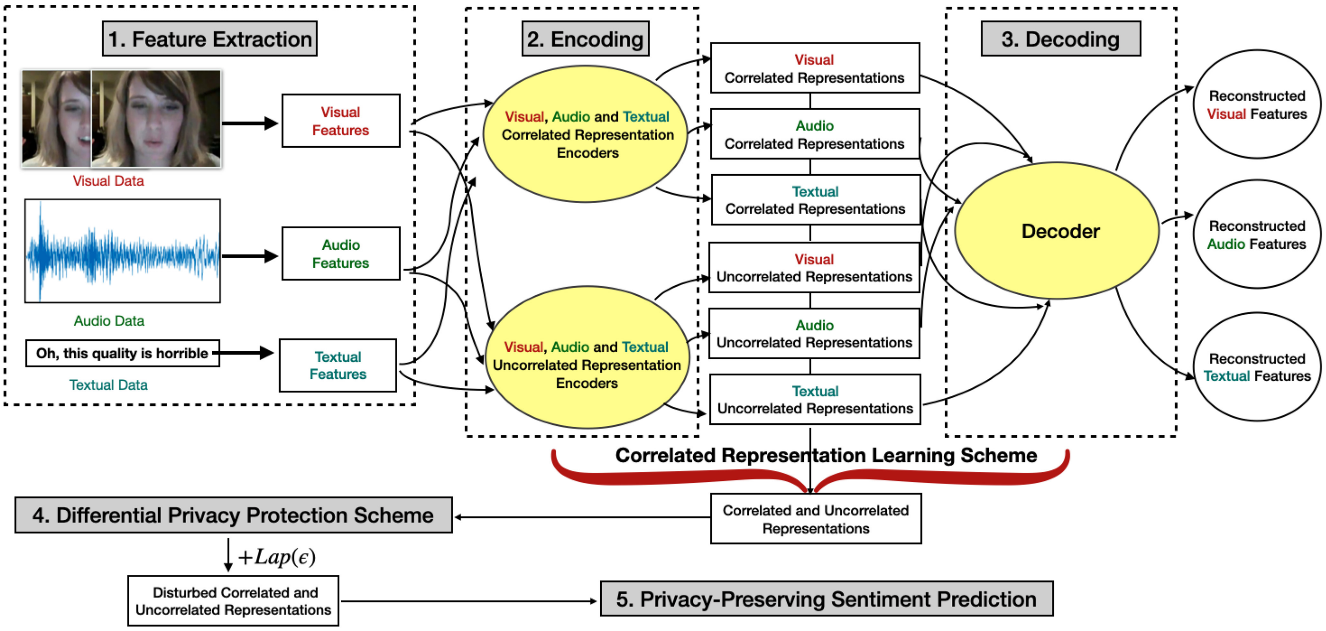
1. Feature Extraction
Visual Data
Audio Data
Oh, this quality is horrible
Textual Data
Visual Features
Audio Features
Textual Features
2. Encoding
Visual, Audio and Textual Correlated Representation Encoders
Visual, Audio and Textual Uncorrelated Representation Encoders
Visual Correlated Representations
Audio Correlated Representations
Textual Correlated Representations
Visual Uncorrelated Representations
Audio Uncorrelated Representations
Textual Uncorrelated Representations
3. Decoding
Decoder
Reconstructed Visual Features
Reconstructed Audio Features
Reconstructed Textual Features
Correlated Representation Learning Scheme
Correlated and Uncorrelated Representations
4. Differential Privacy Protection Scheme
+Lap(ϵ)
Disturbed Correlated and Uncorrelated Representations
5. Privacy-Preserving Sentiment Prediction

Fig. 2. Data flow of our DPCRL model.

multimodal representation encoder E_m^c to extract the corresponding correlated representation $\mathbf{f}_m^c \in \mathbb{R}^{d_h}$ and employ the uncorrelated multimodal representation encoder E_m^u to capture the corresponding uncorrelated representation $\mathbf{f}_m^u \in \mathbb{R}^{d_h}$

$$\mathbf{f}_m^c = E_m^c(\mathbf{f}_m; \theta_m^c, c) \tag{2}$$

$$\mathbf{f}_m^u = E_m^u(\mathbf{f}_m; \theta_m^u) \tag{3}$$

where θ_m^c represents the parameters of the encoder E_m^c, θ_m^u represents the parameters of the encoder E_m^u, and c represents an expected correlation factor that is set to obtain the correlated representations with the expected correlation.

The orthogonality constraint can be used to achieve nonredundancy between two representations. Therefore, to satisfy the first and the second requirements of encoding, we formulate the *data orthogonality loss*, $\mathcal{L}_{\text{enc}_1}$

$$\mathcal{L}_{\text{enc}_1} = \sum_{m\in\{v,a,l\}} ||\mathbf{f}_m^{c\,T}\mathbf{f}_m^u||_F^2 + \sum_{m\neq m'\in\{v,a,l\}} ||\mathbf{f}_m^{u\,T}\mathbf{f}_{m'}^u||_F^2 \tag{4}$$

where $||\cdot||_F^2$ is the squared Frobenius norm.

Then, inspired by the idea of [66], we use the cosine distance to quantify the correlation between two correlated representations. Considering the third requirement of encoding, we define the *data correlation loss*, $\mathcal{L}_{\text{enc}_2}$

$$\mathcal{L}_{\text{enc}_2} = \sum_{m\neq m'\in\{v,a,l\}} ||\mathbf{f}_m^{c\,T}\mathbf{f}_{m'}^c - cI||_F^2 \tag{5}$$

where $c \in [0, 1]$ is a correlation factor that indicates the cosine distance between two representations, and I denotes the identity matrix. To sum up, the entire encoding loss function $\mathcal{L}_{\text{enc}}$ is the summation of $\mathcal{L}_{\text{enc}_1}$ in (4) and $\mathcal{L}_{\text{enc}_2}$ in (5), shown in

$$\mathcal{L}_{\text{enc}} = \mathcal{L}_{\text{enc}_1} + \mathcal{L}_{\text{enc}_2}. \tag{6}$$

C. Decoding

Since an encoder function may output an unrepresentative vector that cannot be recovered, we design a decoder D to reconstruct the original feature vector by using the extracted correlated and uncorrelated representations (*i.e.*, $\mathbf{f}_m^c$ and $\mathbf{f}_m^u$) in each modality. The decoder D is defined in (7) to ensure that the encoded representations indeed represent the details of the corresponding modality data [4], [49]

$$\bar{\mathbf{f}}_m = D(\mathbf{f}_m^c + \mathbf{f}_m^u; \theta_d) \tag{7}$$

where $\bar{\mathbf{f}}_m$ is the reconstructed feature vector for $m \in \{v, a, l\}$, and θ_d represents the parameters of the decoder D. In the decoding process, the reconstruction loss, $\mathcal{L}_{\text{dec}}$, is measured by *mean-squared error* as below

$$\mathcal{L}_{\text{dec}} = \sum_{m\in\{v,a,l\}} \frac{||\mathbf{f}_m - \bar{\mathbf{f}}_m||_2^2}{d_h} \tag{8}$$

where $||\cdot||_2^2$ denotes the squared $L2$-norm.

Finally, the correlated representation learning can be achieved through the autoencoding architecture that is the combination of the encoders and the decoders. Correspondingly, the loss function of the correlated representation learning process, $\mathcal{L}_{CRL}$, is the summation of the encoding loss $\mathcal{L}_{\text{enc}}$ in (6), and the decoding loss $\mathcal{L}_{\text{dec}}$ in (8), i.e.,

$$\mathcal{L}_{CRL} = \alpha\mathcal{L}_{\text{enc}} + \beta\mathcal{L}_{\text{dec}} \tag{9}$$

where $\alpha \in (0, 1]$ and $\beta \in (0, 1]$ are the weights of loss functions. We minimize $\mathcal{L}_{CRL}$ to obtain the correlated and uncorrelated multimodal representations for multimodal sentiment analysis.

D. Differential Privacy Protection Scheme

After obtaining the correlated and uncorrelated representations through our proposed correlated representation learning, we implement the differential privacy mechanisms to generate privacy-preserving representations for multimodal sentiment analysis. To be specific, in our differential privacy protection scheme, the representations captured by our proposed correlated representation learning and the privacy-preserving representations are considered as the neighboring databases in the differential privacy theory. In the following, we apply different differential privacy mechanisms to the correlated and uncorrelated representations.

First, according to *basic differential privacy mechanism* [67], we can calculate the perturbed uncorrelated representation $\hat{\mathbf{f}}^u_m = \mathbf{f}^u_m + \mathrm{Lap}(0, S_{\mathbf{f}^u_m}/\epsilon)$ by using an additional Laplace noise to satisfy ϵ-differential privacy, where $S_{\mathbf{f}^u_m}$ represents the global sensitivity of the uncorrelated representation vector $\mathbf{f}^u_m$ and is equal to the difference between the maximal and the minimal items in $\mathbf{f}^u_m$.

Theorem 1: Given the Laplace noise $\mathrm{Lap}(0, S_{\mathbf{f}^u_m}/\epsilon)$ added into the uncorrelated representation vector $\mathbf{f}^u_m$, the disturbed uncorrelated representation vector $\hat{\mathbf{f}}^u_m$ satisfies ϵ-differential privacy.

Proof: Let $\Pr[\,\cdot\,]$ be a commonly designed Laplace distribution [68]. Accordingly, we have

$$\begin{aligned}\ln\frac{\Pr[\mathbf{f}^u_m]}{\Pr[\hat{\mathbf{f}}^u_m]} &= \ln\frac{\frac{\epsilon}{2S_{\mathbf{f}^u_m}}e^{-\frac{\epsilon}{S_{\mathbf{f}^u_m}}|\mathbf{f}^u_m|}}{\frac{\epsilon}{2S_{\mathbf{f}^u_m}}e^{-\frac{\epsilon}{S_{\mathbf{f}^u_m}}|\hat{\mathbf{f}}^u_m|}}\\ &= \frac{\epsilon}{S_{\mathbf{f}^u_m}}(|\hat{\mathbf{f}}^u_m| - |\mathbf{f}^u_m|) \le \epsilon. \end{aligned}\tag{10}$$

Equation (10) shows that the disturbed uncorrelated representation vector $\hat{\mathbf{f}}^u_m$ satisfies ϵ-differential privacy. ■

Second, we use *correlated differential privacy mechanism* [69] to achieve the correlated representations' ϵ-differential privacy by adding Laplace noise. In this article, we use the non-negative cosine distance $\mathrm{Cos}(\cdot,\cdot) \in [0, 1]$ to measure the correlation among representations, where a higher cosine distance value means a larger correlation, and a lower cosine distance value indicates a smaller correlation. Then, we can compute the perturbed correlated representation $\hat{\mathbf{f}}^c_m$ as

$$\hat{\mathbf{f}}^c_m = \mathbf{f}^c_m + \mathrm{Lap}\left(0, \sum_{m'\in\{v,a,l\}}\mathrm{Cos}(\mathbf{f}^c_m, \mathbf{f}^c_{m'})S_{\mathbf{f}^c_m}/\epsilon\right)\tag{11}$$

where $S_{\mathbf{f}^c_m}$ is the global sensitivity of the uncorrelated representation vector $\mathbf{f}^c_m$ and is equal to the difference between the maximal and the minimal items in $\mathbf{f}^c_m$, and $\mathrm{Cos}(\mathbf{f}^c_m, \mathbf{f}^c_{m'})$ is used as the correlation coefficient between $\mathbf{f}^c_m$ and $\mathbf{f}^c_{m'}$.

Theorem 2: By adding the Laplace noise $\mathrm{Lap}(0, \sum_{m'\in\{v,a,l\}}\mathrm{Cos}(\mathbf{f}^c_m, \mathbf{f}^c_{m'})S_{\mathbf{f}^c_m}/\epsilon)$ into the correlated representation vector $\mathbf{f}^c_m$, the output perturbed correlated representation vector $\hat{\mathbf{f}}^c_m$ meets ϵ-differential privacy.

Proof: In accordance with [69], we define $QS_{\mathbf{f}^c_m} = \sum_{m'\in\{v,a,l\}}\mathrm{Cos}(\mathbf{f}^c_m, \mathbf{f}^c_{m'})S_{\mathbf{f}^c_m}$ as the correlated global sensitivity of the correlated representation vector $\mathbf{f}^c_m$. Similar to the proof of Theorem 1, let $\Pr[\,\cdot\,]$ be the Laplace distribution. Accordingly, there is

$$\begin{aligned}\ln\frac{\Pr[\mathbf{f}^c_m]}{\Pr[\hat{\mathbf{f}}^c_m]} &= \ln\frac{\frac{\epsilon}{2QS_{\mathbf{f}^c_m}}e^{-\frac{\epsilon}{QS_{\mathbf{f}^c_m}}|\mathbf{f}^c_m|}}{\frac{\epsilon}{2QS_{\mathbf{f}^c_m}}e^{-\frac{\epsilon}{QS_{\mathbf{f}^c_m}}|\hat{\mathbf{f}}^c_m|}}\\ &= \frac{\epsilon}{QS_{\mathbf{f}^c_m}}(|\hat{\mathbf{f}}^c_m| - |\mathbf{f}^c_m|) \le \epsilon. \end{aligned}\tag{12}$$

Equation (12) indicates that the perturbed correlated representation vector $\hat{\mathbf{f}}^c_m$ meets ϵ-differential privacy. ■

Notably, for $\hat{\mathbf{f}}^c_m$, the added Laplace noise can be lower if the value of $\mathrm{Cos}(\mathbf{f}^c_m, \mathbf{f}^c_{m'})$ is decreased, which can mitigate the side-effect of the Laplace noise on the sentiment prediction performance. On the other hand, as shown in $\mathcal{L}_{\mathrm{enc}_2}$, the correlation between $\mathbf{f}^c_m$ and $\mathbf{f}^c_{m'}$ can be adjusted by changing the value of c in our correlated representation learning process, which makes the generation of privacy-preserving representations more flexible.

E. Privacy-Preserving Sentiment Prediction

Following the fusion idea of [49], the outputs of the aforementioned differential privacy protection scheme, including $\hat{\mathbf{f}}^c_v$, $\hat{\mathbf{f}}^c_a$, $\hat{\mathbf{f}}^c_l$, $\hat{\mathbf{f}}^u_v$, $\hat{\mathbf{f}}^u_a$, and $\hat{\mathbf{f}}^u_l$, are fused into a joint vector $\hat{\mathbf{f}}_{\mathrm{out}} \in \mathbb{R}^{d_{\mathrm{out}}}$ through simple concatenation. Then, the prediction function G is applied to the privacy-preserving prediction task with $\hat{\mathbf{f}}_{\mathrm{out}}$ as the input

$$\hat{\mathbf{y}} = G(\hat{\mathbf{f}}_{\mathrm{out}}; \theta_{\mathrm{out}})\tag{13}$$

where $\hat{\mathbf{y}}$ is the predicted label vector corresponding to $\hat{\mathbf{f}}_{\mathrm{out}}$, and θ_{out} denotes the parameters of the prediction function.

We use *cross-entropy loss* to calculate the loss of the privacy-preserving sentiment prediction task in

$$\mathcal{L}_{\mathrm{task}} = -\frac{1}{n}\sum_{i=0}^{n}\mathbf{y}_i \cdot \log(\hat{\mathbf{y}}_i)\tag{14}$$

in which $\mathcal{L}_{\mathrm{task}}$ is the prediction loss, n represents the number of utterances in a training batch, $\mathbf{y}_i$ is the ith ground-truth label, and $\hat{\mathbf{y}}_i$ is the ith predicted label.

Consequently, to learn the privacy-preserving correlated and uncorrelated multimodal representations for the privacy-preserving multimodal sentiment analysis, the overall loss function of DPCRL, $\mathcal{L}_{\mathrm{DPCRL}}$, should consist of the encoding loss $\mathcal{L}_{\mathrm{enc}}$ in (6), the decoding loss $\mathcal{L}_{\mathrm{dec}}$ in (7), and the privacy-preserving prediction loss $\mathcal{L}_{\mathrm{task}}$ in (14) as formulated by

$$\mathcal{L}_{\mathrm{DPCRL}} = \alpha\mathcal{L}_{\mathrm{enc}} + \beta\mathcal{L}_{\mathrm{dec}} + \gamma\mathcal{L}_{\mathrm{task}}\tag{15}$$

where $\alpha, \beta, \gamma \in (0, 1]$ are the weights of the loss functions. Our DPCRL model can be learnt by minimizing $\mathcal{L}_{\mathrm{DPCRL}}$. The specific network architectures of the encoders, E^c_m and E^u_m, the decoder D, and the prediction function G used in the DPCRL model are described in Section IV-A4.

IV. Experiments

In this section, we first introduce our experiment settings and then present comprehensive experimental results to validate the superiority of our proposed DPCRL model

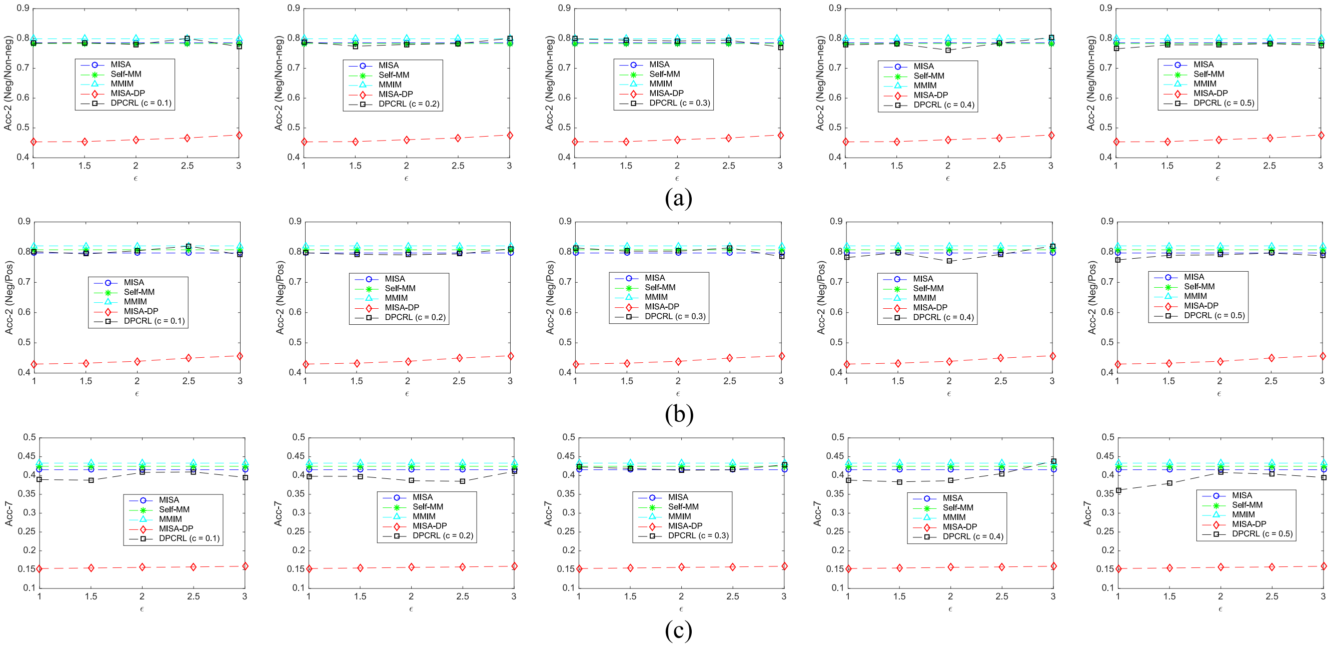

Fig. 3. Comparison results of DPCRL with different c on MOSI dataset (versus baselines). (a) Evaluation results of Acc-2 (Neg/Non-neg) on MOSI dataset (DPCRL with different c and ϵ and baselines). (b) Evaluation results of Acc-2 (Neg/Pos) on MOSI dataset (DPCRL with different c and ϵ and baselines). (c) Evaluation results of Acc-7 on MOSI dataset (DPCRL with different c and ϵ and baselines).

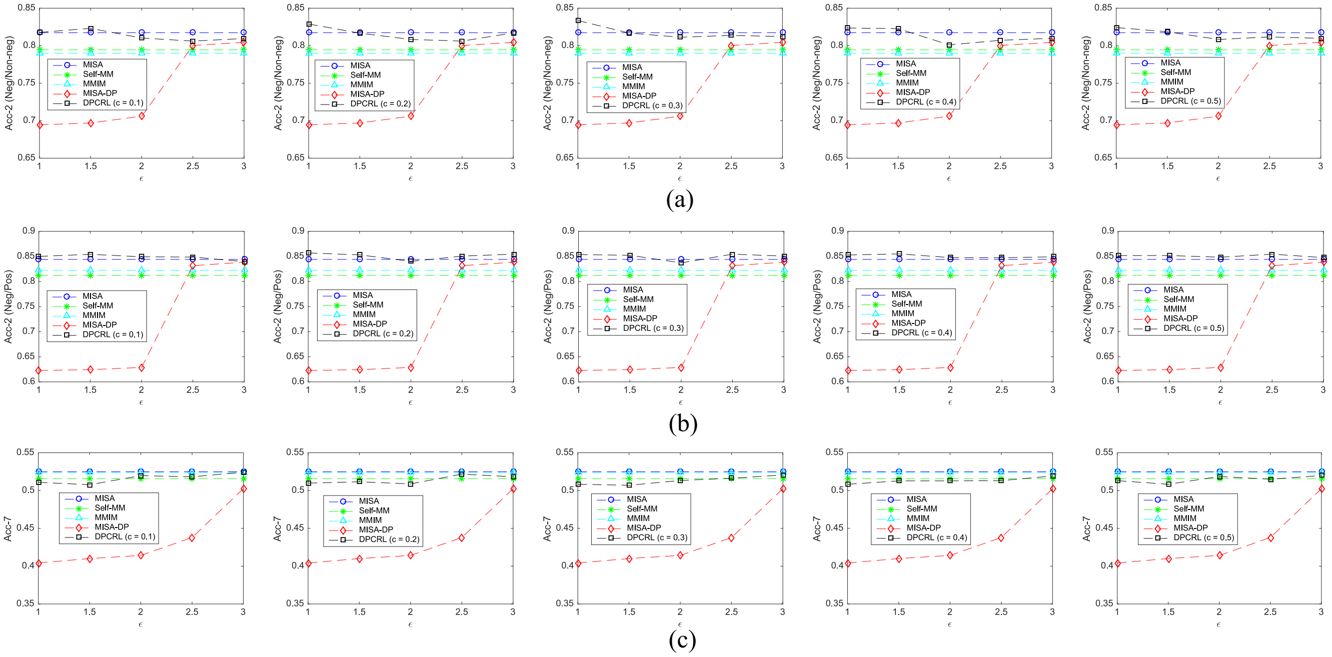

Fig. 4. Comparison results of DPCRL with different c on MOSEI dataset (versus baselines). (a) Evaluation results of Acc-2 (Neg/Non-neg) on MOSEI dataset (DPCRL with different c and ϵ and baselines). (b) Evaluation results of Acc-2 (Neg/Pos) on MOSEI dataset (DPCRL with different c and ϵ and baselines). (c) Evaluation results of Acc-7 on MOSEI dataset (DPCRL with different c and ϵ and baselines).

over the state-of-the-art for privacy-preserving multimodal sentiment analysis. The codes of our model and all experimental results in this article can be found at https://github.com/ahahnut/DPCRL-for-Privacy-Preserving-Multimodel-Sentiment-Analysis.

A. Experimental Settings

The datasets, baselines, performance metrics, network architectures, and hyperparameter settings are described below.

1) Datasets: We use two benchmark datasets in our experiments for multimodal sentiment analysis. *CMU-MOSI (MOSI) dataset* [70] is a collection of YouTube monologues consisting of 2198 subjective video segments (utterances), where speakers express their opinions on topics, such as movies. Each utterance is manually annotated with an integer opinion score in $[-3, 3]$, where -3 and 3 represent the strongest negative and the strongest positive sentiments, respectively. *CMU-MOSEI (MOSEI) dataset* [50] contains 23 453 annotated video segments and is an improvement of MOSI with a larger number of utterances and a greater variety in samples, speakers, and topics.

2) Baseline: MISA [49], Self-MM [71], and MMIM [72] are the currently pioneering models on both MOSI and MOSEI datasets for multimodal sentiment analysis. MISA with differential privacy (MISA-DP) is a simple combination of the differential privacy mechanism and MISA to obtain differentially private representations for sentiment prediction while guaranteeing privacy protection. MISA, Self-MM, MMIM, and MISA-DP are adopted as baseline mechanisms for performance comparison.

3) Performance Metrics: The task of sentiment prediction on MOSI and MOSEI can be treated as a classification process and evaluated via integer classification scores in $[-3, 3]$ that are so-called seven-class accuracy (Acc-7) [70]. Besides, two approaches of computing binary accuracy (Acc-2) can be also adopted to measure the performance of sentiment prediction. The first one is *negative/non-negative* (*Neg/Non-neg*) classification, where the non-negative labels are indicated by non-negative classification scores [53]. The second one is calculated based on *negative/positive* (*Neg/Pos*) classes, where the negative and the positive classes are indicated by the negative and the positive scores, respectively [73]. To sum up, Acc-2 (Neg/Non-neg), F1 (Neg/Non-neg), Acc-2 (Neg/Pos), F1 (Neg/Pos), and Acc-7 are used as performance metrics in our experiments.

4) Neural Network Architectures: In our proposed DPCRL model, the neural network architectures of the feature extraction, encoding, decoding, and sentiment prediction modules are described below.

1) *Feature Extraction:* Facial action coding system (FACS) [74] is applied to extract facial expression features that include facial action units and face pose. An acoustic analysis framework (COVAREP) [75] is employed to extract the acoustic features that contain 12 Mel-frequency cepstral coefficients, pitch, voiced/unvoiced segmenting features, glottal source parameters, and other features related to emotions and the tone of speech. The pretrained BERT [76] is utilized as the feature extractor for textual utterance. Accordingly, the visual feature dimension is $d_v = 47$, the acoustic feature dimension is $d_a = 74$, and the textual feature dimension is $d_l = 784$. Furthermore, in order to align the multimodal features for our encoding process, we exploit *one fully connected layer with ReLU activation function and one normalization layer* to embed these features into a space with the same dimension.
2) *Encoding:* The correlated multimodal representation encoder E_m^c is built by using *one fully connected layer with sigmoid activation function* to extract the correlated representations. The uncorrelated multimodal representation encoder E_m^u is designed through *one fully connected layer with sigmoid activation function* to extract the uncorrelated representations. To be specific, there are three encoders to learn the correlated representations and three encoders to learn the uncorrelated representations. Although these encoders have the same structure, their parameters are updated differently during training process to learn correlated and uncorrelated representations.
3) *Decoding:* The decoder D is established as *one fully connected layer* for reconstruction to avoid learning unrepresentative vector of data in the encoding process.
4) *Sentiment Prediction:* In the prediction function G, *one Transformer encoder layer* is used for transformation, *one fully connected layer with a dropout layer plus a ReLU activation function* is used for fusion, and *one fully connected layer* is used to map all representations into one dimension for final prediction.

5) Hyperparameter Settings: Our experiments are conducted on Ubuntu OS with a Nvidia Tesla V100 GPU and 16 GB RAM. The batch size of samples for training MOSI and MOSEI datasets are 64 and 16, respectively. The learning rate of training is set as 10^{-4}. The probabilities of dropout in the dropout layer for training MOSI and MOSEI datasets are 0.5 and 0.1, respectively. Via comprehensive ablation study, the weights of loss functions are set as $\alpha = 0.45$, $\beta = 0.1$, and $\gamma = 0.45$ for training the MOSI dataset with 500 epochs, and the weights of loss functions are set to be $\alpha = 0.35$, $\beta = 0.3$, and $\gamma = 0.35$ for training the MOSEI dataset with 500 epochs. Besides, we vary the correlation factor c from 0 to 1 with the step of 0.1 to illustrate the effectiveness of our correlated representation learning model and set the privacy budget $\epsilon \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$ to evaluate our DPCRL model.

B. Evaluation on Our DPCRL Model

In our proposed DPCRL model, there are two system parameters, ϵ and c. The value of ϵ, which is so-called "privacy budget," indicates the degree of privacy protection. A smaller ϵ implies a higher degree of data privacy protection. We implement our DPCRL model with $\epsilon = 1.0, 1.5, 2.0, 2.5, 3.0$ on datasets, which is reasonable and applicable in real applications for privacy protection based on the differential privacy mechanisms. The value of c represents the expected correlation among the learned correlated representations. A larger c implies a closer correlation among the correlated representations. In our experiments, we set $c = 0.1, 0.2, 0.3, 0.4, 0.5$ with the following considerations.

1) From Tables II and III, the prediction performance of our correlated representation learning scheme with $c = 0.0$ is worse than that of the state-of-the-art (MISA). Therefore, it may not be suitable to set $c = 0.0$ when we aim to maintain prediction performance as much as possible while ensuring differential privacy protection.
2) We attempt to learn the correlated representations with a relatively lower value of c so as to decrease the side-effect of the additional Laplace noise on prediction performance.

In Fig. 3, we compare the Acc-2 (Neg/Non-neg) results of our DPCRL model and the baseline models on the MOSI dataset. We take Acc-2 (Neg/Non-neg) of DPCRL with $c = 0.1$ as an example to illustrate the effectiveness of our proposed DPCRL model as follows.

1) By comparing the Acc-2 (Neg/Non-neg) values, it can be found that the performance of DPCRL is comparable to that of baselines (including MISA, Self-MM, and MMIM), which indicates that our DPCRL model can maintain the performance of sentiment analysis while satisfying differential privacy guarantee.
2) By comparing Acc-2 (Neg/Non-neg) values of MISA-DP and DPCRL with a same value of ϵ, we can see that the Acc-2 (Neg/Non-neg) values of our proposed DPCRL model are much higher than those of the baseline model MISA-DP which uses the invariant data representations with the correlation $c = 1.0$.

That is, with the same privacy budget ϵ, our DPCRL model outperforms MISA-DP from the aspect of maintaining the sentiment prediction performance. The main reason is that our correlated representation learning scheme used in DPCRL can be leveraged to learn the correlated representations with a relatively lower correlation factor, mitigating the side-effect of the additional Laplace noise on the sentiment analysis.

For a comprehensive demonstration, we present Acc-2 (Neg/Pos), F1 (Neg/Non-neg, Neg/Pos), and Acc-7 of our DPCRL model and the baselines on the MOSI dataset in Fig. 3. Additionally, for the MOSEI dataset, the values of Acc-2 (Neg/Non-neg, Neg/Pos), F1 (Neg/Non-neg, Neg/Pos), and Acc-7 of our DPCRL model and baselines are presented in Fig. 4.

Based on the above analysis, we obtain the following critical conclusions.

1) Our proposed DPCRL model is effective to accomplish privacy-preserving multimodal sentiment analysis with providing ϵ-differential privacy guarantee.
2) By setting a correlation factor as input, our DPCRL model can realize heterogeneous multimodal data transformation that satisfies our learning expectation.
3) A smaller value of the correlation factor can help reduce Laplace noise added in ϵ-differential privacy mechanisms, mitigating the loss of prediction performance.
4) Compared with the state-of-the-art, our DPCRL model can effectively maintain and even enhance the performance of sentiment prediction while ensuring ϵ-differential privacy.

Furthermore, more experiments are conducted using an NVIDIA V100-16GB GPU and an AMD 64-Core CPU. This hardware configuration ensures that both our DPCRL model and the baselines are operated under the same conditions for a fair comparative analysis. We evaluate the computational cost based on running time in each training epoch, memory usage, and CPU/GPU utilization. These metrics are critical for assessing the impact of differential privacy mechanisms on model efficiency and resource consumption. The results are summarized in Table I, which indicates that our DPCRL incurs a 16.7% increase in running time compared to the fastest baseline (MMIM), demonstrating the additional time required for processing privacy-preserving mechanisms. The memory usage in our proposed DPCRL is 8.3% higher than the least memory-intensive baseline (MISA), which reflects the extra memory is required for handling DP operations. Compared to the baseline with the lowest CPU/GPU utilization (MISA), the increase in CPU utilization is 11.4% and the increase in GPU utilization is 12.5%.

TABLE I
COMPUTATIONAL COST COMPARISON AMONG DPCRL AND BASELINES

Metric	MISA (Non-DP)	Self-MM (Non-DP)	MMIM (Non-DP)	DPCRL	% Increase over Best Baseline
Running Time (s)	95 s	100 s	**90 s**	105 s	16.7% (vs MMIM)
Memory Usage (GB)	**9.6 GB**	10.2 GB	10 GB	10.4 GB	8.3% (vs MISA)
CPU Utilization (%)	**70%**	76%	75%	78%	11.4% (vs MISA)
GPU Utilization (%)	**64%**	70%	68%	72%	12.5% (vs MISA)

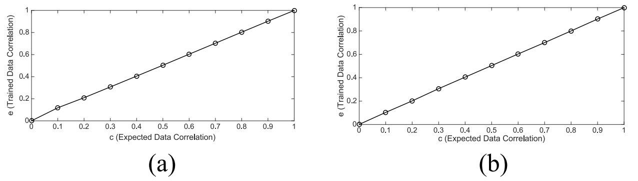

Fig. 5. Impact of expected data correlation c on trained data correlation e. (a) Trained data correlation in MOSI dataset. (b) Trained data correlation in MOSEI dataset.

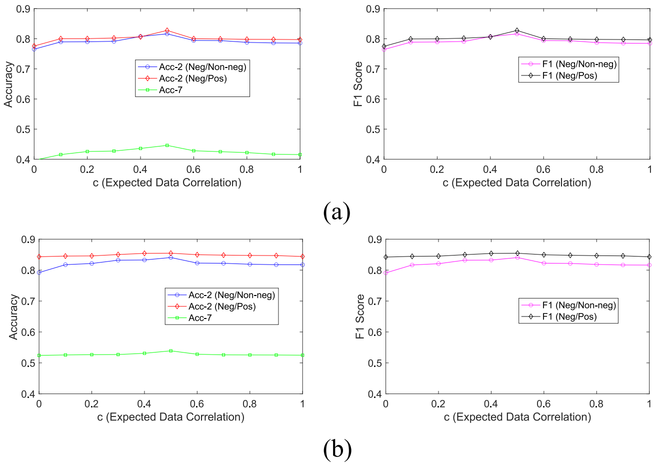

Fig. 6. Impact of expected data correlation c on prediction results of CRL. (a) Prediction results of CRL on MOSI dataset. (b) Prediction results of CRL on MOSEI dataset.

C. Ablation Study

We first train our scheme by changing the correlation factor c from 0 to 1 with the step of 0.1 to validate that c can help achieve effective heterogeneous multimodal data transformation satisfying the requirements for multimodal sentiment analysis. When the training process terminates, the correlation coefficient among the trained correlated representations is denoted by e. Since $c, e \in [0, 1]$ are the cosine values, we can calculate the angle degree, d_c, corresponding to c and the angle degree, d_e, corresponding to e. That is, c and d_c imply our expected data correlation, and e and d_e are our trained data correlation. The difference between our expected and trained data correlation can reflect the effectiveness of our proposed correlated representation learning scheme. To clearly investigate the impact of c on the performance of sentiment prediction, we compute Acc-2 (Neg/Non-neg), F1 (Neg/Non-neg), Acc-2 (Neg/Pos), F1 (Neg/Pos), and Acc-7 on the learned correlated and uncorrelated representations.

Table II presents the values of c, d_c, e, and d_e when the correlated representation learning scheme is implemented on

TABLE II
EVALUATION RESULTS OF CORRELATED REPRESENTATION LEARNING SCHEME ON MOSI DATASET

Model	Expected Data Correlation	Trained Data Correlation	Acc-2 (Neg/Non-neg, Neg/Pos)	F1 (Neg/Non-neg, Neg/Pos)	Acc-7
MISA [49]	/	/	0.7857/0.7972	0.7847/0.8092	0.4154
Self-MM [71]	/	/	0.783/0.8079	0.7834/0.8066	0.4244
MMIM [72]	/	/	0.799/0.8208	0.7984/0.8173	0.433
CRL	$c = 0.0,\ d_c = 90.00°$	$e = 0.0003,\ d_e = 89.82°$	0.7653/0.7759	0.7643/0.7749	0.3965
CRL	$c = 0.1,\ d_c = 84.26°$	$e = 0.1183,\ d_e = 83.21°$	0.7896/0.8003	0.7887/0.7994	0.4154
CRL	$c = 0.2,\ d_c = 78.46°$	$e = 0.2093,\ d_e = 77.92°$	0.79/0.8004	0.7893/0.7997	0.4256
CRL	$c = 0.3,\ d_c = 72.54°$	$e = 0.3069,\ d_e = 72.13°$	0.7915/0.8024	0.791/0.8019	0.4271
CRL	$c = 0.4,\ d_c = 66.42°$	$e = 0.4050,\ d_e = 66.11°$	0.8075/0.8064	0.8072/0.8061	0.4358
CRL	$c = 0.5,\ d_c = 60.00°$	$e = 0.5036,\ d_e = 59.76°$	**0.8163/0.8277**	**0.8162/0.8276**	**0.446**
CRL	$c = 0.6,\ d_c = 53.13°$	$e = 0.6035,\ d_e = 52.88°$	0.7944/0.8	0.7941/0.8003	0.4281
CRL	$c = 0.7,\ d_c = 45.57°$	$e = 0.7029,\ d_e = 45.34°$	0.794/0.7994	0.7935/0.7989	0.425
CRL	$c = 0.8,\ d_c = 36.87°$	$e = 0.8029,\ d_e = 36.59°$	0.788/0.7982	0.7873/0.7979	0.422
CRL	$c = 0.9,\ d_c = 25.84°$	$e = 0.9023,\ d_e = 25.54°$	0.7862/0.7979	0.7853/0.7974	0.4165
CRL	$c = 1.0,\ d_c = 0.00°$	$e = 0.9997,\ d_e = 1.40°$	0.7857/0.7972	0.7847/0.7962	0.4154

TABLE III
EVALUATION RESULTS OF CORRELATED REPRESENTATION LEARNING SCHEME ON MOSEI DATASET

Model	Expected Data Correlation	Trained Data Correlation	Acc-2 (Neg/Non-neg, Neg/Pos)	F1 (Neg/Non-neg, Neg/Pos)	Acc-7
MISA [49]	/	/	0.8173/0.844	0.8193/0.841	0.5249
Self-MM [71]	/	/	0.7944/0.8122	0.7995/0.825	0.5159
MMIM [72]	/	/	0.79/0.8223	0.7966/0.8351	0.5237
CRL	$c = 0.0,\ d_c = 90.00°$	$e = 0.0007,\ d_e = 89.60°$	0.792/0.8432	0.791/0.8422	0.5242
CRL	$c = 0.1,\ d_c = 84.26°$	$e = 0.1034,\ d_e = 84.07°$	0.8175/0.8454	0.8166/0.8445	0.5257
CRL	$c = 0.2,\ d_c = 78.46°$	$e = 0.2017,\ d_e = 78.36°$	0.8214/0.8459	0.8207/0.8452	0.5266
CRL	$c = 0.3,\ d_c = 72.54°$	$e = 0.3066,\ d_e = 72.15°$	0.8321/0.8503	0.8321/0.8498	0.527
CRL	$c = 0.4,\ d_c = 66.42°$	$e = 0.4052,\ d_e = 66.10°$	0.8327/0.8542	0.8324/0.8539	0.531
CRL	$c = 0.5,\ d_c = 60.00°$	$e = 0.5040,\ d_e = 59.74°$	**0.8407/0.8547**	**0.8406/0.8546**	**0.539**
CRL	$c = 0.6,\ d_c = 53.13°$	$e = 0.6034,\ d_e = 52.89°$	0.8227/0.8498	0.8224/0.8495	0.528
CRL	$c = 0.7,\ d_c = 45.57°$	$e = 0.7005,\ d_e = 45.53°$	0.8221/0.8484	0.8216/0.8479	0.526
CRL	$c = 0.8,\ d_c = 36.87°$	$e = 0.8004,\ d_e = 36.83°$	0.819/0.8474	0.8183/0.8467	0.5257
CRL	$c = 0.9,\ d_c = 25.84°$	$e = 0.9022,\ d_e = 25.55°$	0.8175/0.847	0.8166/0.8461	0.5255
CRL	$c = 1.0,\ d_c = 0.00°$	$e = 0.9996,\ d_e = 1.62°$	0.8173/0.844	0.8163/0.843	0.5249

the MOSI dataset. By comparing these values, one can see that the expected data correlation is very close to the corresponding trained data correlation. For examples, $e = 0.1183$ when $c = 0.1$, and $e = 0.2093$ when $c = 0.2$. For a more explicit comparison, we plot Fig. 5(a), to examine the impact of c on e, from which we can also observe that e is nearly equal to c. The results of Table II and Fig. 5(a), confirm that in our correlated representation learning scheme, the utilization of c is effective to accomplish our expected heterogeneous multimodal data transformation. When implementing our correlated representation learning scheme on the MOSEI dataset, we can obtain the same conclusion through Table III and Fig. 5(b).

Additionally, the multimodal representations learned from our correlated representation learning scheme are exploited to evaluate the performance of sentiment analysis in terms of Acc-2 (Neg/Non-neg), F1 (Neg/Non-neg), Acc-2 (Neg/Pos), F1 (Neg/Pos), and Acc-7. These experimental results on MOSI dataset are presented in Table II. Take the values of Acc-2 (Neg/Non-neg) as an example for analysis as follows.

1) The values of Acc-2 (Neg/Non-neg) obtained via MISA, Self-MM, and MMIM are 0.7857, 0.783, and 0.799, respectively. While, the value of Acc-2 (Neg/Non-neg) obtained in our correlated representation learning scheme falls in $[0.7653, 0.8163]$ when the value of c varies from 0 to 1 with the step of 0.1. Especially, when $c = 0.5$ (*i.e.,* the angle degree is $d_c = 60°$), the value of Acc-2 (Neg/Non-neg) reaches 0.8163. Thus, we can conclude that our correlated representation learning scheme and the baselines (including MISA, Self-MM, and MMIM) have comparable performance in terms of Acc-2 (Neg/Non-neg).
2) For our correlated representation learning scheme, the value of Acc-2 (Neg/Non-neg) increases with the growth of c when $c \in [0.0, 0.5]$, which indicates that the increased similarity among representations is helpful to improve the performance of sentiment prediction.
3) In our correlated representation learning scheme, the value of Acc-2 (Neg/Non-neg) gradually decreases with the growth of c when $c \in [0.6, 1.0]$, which implies that the decreased diversity among representations degrades the performance of sentiment prediction.
4) The correlation factor c can be used to balance the trade-off between representation similarity and representation diversity for improving multimodal sentiment analysis performance.

Similarly, by analyzing the results of F1 (Neg/Non-neg), Acc-2 (Neg/Pos), F1 (Neg/Pos), and Acc-7 on the MOSI dataset in Table II, we can draw the same conclusions. In order to explicitly show the impact of c on sentiment prediction, we present the results of Acc-2 (Neg/Non-neg), F1 (Neg/Non-neg), Acc-2 (Neg/Pos), F1 (Neg/Pos), and Acc-7 on the MOSI dataset in Fig. 6(a), for comparison. Moreover, as shown in

TABLE IV
ABLATION STUDY OF CORRELATED REPRESENTATION LEARNING SCHEME ON MOSI DATASET

Uncorrelated Representations	Correlated Representations	Acc-2 (Neg/Non-neg, Neg/Pos)	F1 (Neg/Non-neg, Neg/Pos)	Acc-7
✓	×	0.6407/0.6461	0.6547/0.6527	0.3145
×	✓	0.7084/0.7008	0.7027/0.7034	0.3474
✓	✓	**0.8163/0.8277**	**0.8162/0.8276**	**0.446**

TABLE V
ABLATION STUDY OF CORRELATED REPRESENTATION LEARNING SCHEME ON MOSEI DATASET

Uncorrelated Representations	Correlated Representations	Acc-2 (Neg/Non-neg, Neg/Pos)	F1 (Neg/Non-neg, Neg/Pos)	Acc-7
✓	×	0.7053/0.7684	0.7148/0.6827	0.4056
×	✓	0.7832/0.8045	0.7796/0.7608	0.4362
✓	✓	**0.8407/0.8547**	**0.8406/0.8546**	**0.539**

Table III and Fig. 6(b), the experimental results on the MOSEI dataset can also confirm our aforementioned analysis.

Then, we present more ablation study of our correlated representation learning model trained with the correlation factor $c = 0.5$ and the default hyperparameter settings. In Tables IV and V, we show the results of ablation study on the MOSI dataset and the MOSEI dataset, respectively. By comparing these results, it is clear that the incorporation of the correlated and uncorrelated multimodal representations can obtain the best performance of the multimodal sentiment analysis, which verifies the effectiveness of our model design.

V. CONCLUSION

In this article, we proposed a DPCRL model designed to address privacy-preserving challenges in multimodal sentiment analysis, particularly within IoT scenarios. The DPCRL model integrates a novel correlated representation learning scheme with a differential privacy protection scheme, making it suitable for IoT-driven applications, such as smart assistants, healthcare monitoring, and intelligent transportation systems. Our DPCRL model consists of a novel correlated representation learning scheme and a differential privacy protection scheme. The correlated representation learning scheme can achieve heterogeneous multimodal data transformation to learn correlated and uncorrelated representations for multimodal sentiment prediction while reducing privacy leakage. The differential privacy protection scheme can produce the perturbed correlated and uncorrelated representations through inserting Laplace noise for ϵ-differential privacy. In our DPCRL model, a correlation factor is employed to learn the correlated representations for mitigating the side-effect of the additional Laplace noise on the sentiment prediction performance. Finally, the experiment results can confirm that our proposed DPCRL model outperforms the state-of-the-art in the performance of multimodal sentiment prediction and data privacy protection.

REFERENCES

[1] G. Sprint, D. J. Cook, M. Schmitter-Edgecombe, and L. B. Holder, “Multimodal fusion of smart home and text-based behavior markers for clinical assessment prediction,” *ACM Trans. Comput. Healthcare*, vol. 3, no. 4, pp. 1–25, 2022.

[2] Y. Subbarayudu and A. Sureshbabu, “Distributed multimodal aspective on topic model using sentiment analysis for recognition of public health surveillance,” in *Expert Clouds Applications*. Singapore: Springer, 2022, pp. 459–476.

[3] X. Chen, Z. Wang, and X. Di, “Sentiment analysis on multimodal transportation during the COVID-19 using social media data,” *Information*, vol. 14, no. 2, p. 113, 2023.

[4] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[5] S. Khorram, M. Jaiswal, J. Gideon, M. G. McInnis, and E. M. Provost, “The PRIORI emotion dataset: Linking mood to emotion detected in-the-wild,” 2018, *arXiv:1806.10658*.

[6] K. W. Piersol and G. Beddingfield, “Prewakeword speech processing,” U.S. Patent 10 192 546, 2019.

[7] S. Hajian and J. Domingo-Ferrer, “A study on the impact of data anonymization on anti-discrimination,” in *Proc. IEEE 12th Int. Conf. Data Min. Workshops*, 2012, pp. 352–359.

[8] Y. Qu et al., “Toward privacy-aware and trustworthy data sharing using blockchain for edge intelligence,” *Big Data Min. Anal.*, vol. 6, no. 4, pp. 443–464, 2023.

[9] A. Alzu’bi, A. Alomar, S. Alkhaza’leh, A. Abuarqoub, and M. Hammoudeh, “A review of privacy and security of edge computing in smart healthcare systems: Issues, challenges, and research directions,” *Tsinghua Sci. Technol.*, vol. 29, no. 4, pp. 1152–1180, Aug. 2024.

[10] C. Hazman, A. Guezzaz, S. Benkirane, and M. Azrour, “Enhanced IDS with deep learning for IoT-based smart cities security,” *Tsinghua Sci. Technol.*, vol. 29, no. 4, pp. 929–947, Aug. 2024.

[11] Y. Yang, P. Hu, J. Shen, H. Cheng, Z. An, and X. Liu, “Privacy-preserving human activity sensing: A survey,” *High-Confidence Comput.*, vol. 4, Mar. 2024, Art. no. 100204.

[12] L. P. Rachakonda, M. Siddula, and V. Sathya, “A comprehensive study on IoT privacy and security challenges with focus on spectrum sharing in next-generation networks (5G/6G/beyond),” *High-Confidence Comput.*, vol. 4, no. 2, 2024, Art. no. 100220.

[13] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, “A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly,” *High-Confidence Comput.*, vol. 4, no. 2, 2024, Art. no. 100211.

[14] K. Li, G. Luo, Y. Ye, W. Li, S. Ji, and Z. Cai, “Adversarial privacy-preserving graph embedding against inference attack,” *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6904–6915, Apr. 2021.

[15] X. Ding, H. Fang, Z. Zhang, K.-K. R. Choo, and H. Jin, “Privacy-preserving feature extraction via adversarial training,” *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 1967–1979, Apr. 2022.

[16] M. Jaiswal and E. M. Provost, “Privacy enhanced multimodal neural representations for emotion recognition,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 7985–7993.

[17] Z. Xiong, H. Xu, W. Li, and Z. Cai, “Multisource adversarial sample attack on autonomous vehicles,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2822–2835, Mar. 2021.

[18] H. Xu, Z. Cai, D. Takabi, and W. Li, “Audio-visual autoencoding for privacy-preserving video streaming,” *IEEE Internet Things J.*, vol. 9, no. 3, pp. 1749–1761, Feb. 2022.

[19] X. Wang, L. Mo, X. Zheng, and Z. Dang, “Streaming histogram publication over weighted sliding windows under differential privacy,” *Tsinghua Sci. Technol.*, vol. 29, no. 6, pp. 1674–1693, Dec. 2024.

[20] K. Zhang et al., “Toward privacy in decentralized IoT: A blockchain-based dual response DP mechanism,” *Big Data Min. Anal.*, vol. 7, no. 3, pp. 699–717, Sep. 2024.

[21] R. Yan, Y. Zheng, N. Yu, and C. Liang, “Multismart meter data encryption scheme based on distributed differential privacy,” *Big Data Min. Anal.*, vol. 7, no. 1, pp. 131–141, Mar. 2024.

[22] Y. Wang, X. Wu, and D. Hu, “Using randomized response for differential privacy preserving data collection,” in *Proc. EDBT/ICDT Workshops*, 2016, pp. 90–6778.

[23] Y. Wang, X. Wu, and L. Wu, “Differential privacy preserving spectral graph analysis,” in *Proc. Pac.-Asia Conf. Knowl. Discovery Data Min.*, 2013, pp. 329–340.

[24] H. Jiang, S. Sarwar, H. Yu, and S. A. Islam, “Differentially private data publication with multilevel data utility,” *High-Confidence Comput.*, vol. 2, no. 2, 2022, Art. no. 100049.

[25] W. Zhang, G. Yin, Y. Dong, F. Chen, and Q. Zia, “DPTP-LICD: A differential privacy trajectory protection method based on latent interest community detection,” *High-Confidence Comput.*, vol. 3, no. 2, 2023, Art. no. 100134.

[26] Z. Hu and J. Yang, "Differential privacy protection method based on published trajectory cross-correlation constraint," *Plos One*, vol. 15, no. 8, 2020, Art. no. e0237158.

[27] L. Ou, Z. Qin, Y. Liu, H. Yin, Y. Hu, and H. Chen, "Multiuser location correlation protection with differential privacy," in *Proc. IEEE 22nd Int. Conf. Parallel Distrib. Syst.*, 2016, pp. 422–429.

[28] T. Zhang, T. Zhu, P. Xiong, H. Huo, Z. Tari, and W. Zhou, "Correlated differential privacy: Feature selection in machine learning," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2115–2124, Mar. 2020.

[29] H. Wang, Z. Xu, S. Jia, Y. Xia, and X. Zhang, "Why current differential privacy schemes are inapplicable for correlated data publishing?" *World Wide Web*, vol. 24, pp. 1–23, Jan. 2021.

[30] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2010, pp. 735–746.

[31] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 8, pp. 1200–1214, Aug. 2011.

[32] W. Jiang, C. Xie, and Z. Zhang, "Wishart mechanism for differentially private principal components analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1730–1736.

[33] Z. Cai, X. Zheng, J. Wang, and Z. He, "Private data trading toward range counting queries in Internet of Things," *IEEE Trans. Mobile Comput.*, vol. 22, no. 8, pp. 4881–4897, Aug. 2023.

[34] P. Chen, W. He, W. Ma, X. Huang, and C. Wang, "IoTDQ: An industrial IoT data analysis library for Apache IoTDB," *Big Data Min. Anal.*, vol. 7, no. 1, pp. 29–41, Mar. 2024.

[35] Y. Zhao, J. Zhao, and E. Y. Lam, "House price prediction: A multisource data fusion perspective," *Big Data Min. Anal.*, vol. 7, no. 3, pp. 603–620, Sep. 2024.

[36] B. Zhou, J. Liu, S. Cui, and Y. Zhao, "A large-scale spatio-temporal multimodal fusion framework for traffic prediction," *Big Data Min. Anal.*, vol. 7, no. 3, pp. 621–636, Sep. 2024.

[37] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Inf. Fusion*, vol. 91, pp. 424–444, Mar. 2023.

[38] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.

[39] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.

[40] W. Chen and Z. Li, "Octopus v3: Technical report for on-device sub-billion multimodal AI agent," 2024, *arXiv:2404.11459*.

[41] W. Chen, Z. Li, and S. Xin, "OmniVLM: A token-compressed, sub-billion-parameter vision-language model for efficient on-device inference," 2024, *arXiv:2412.11475*.

[42] J. Xu et al., "On-device language models: A comprehensive review," 2024, *arXiv:2409.00088*.

[43] W. Chen, Z. Li, S. Xin, and Y. Wang, "Dolphin: Long context as a new modality for energy-efficient on-device language models," 2024, *arXiv:2408.15518v1*.

[44] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multitask learning for multimodal emotion recognition and sentiment analysis," 2019, *arXiv:1905.05812*.

[45] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual intermodal attention for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 3454–3466.

[46] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2015, pp. 2539–2544.

[47] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016.

[48] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017, *arXiv:1707.07250*.

[49] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[50] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist.*, 2018, pp. 2236–2246.

[51] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Context-aware interactive attention for multimodal sentiment and emotion analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 5651–5661.

[52] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl. Based Syst.*, vol. 161, pp. 124–133, Dec. 2018.

[53] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multiattention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.

[54] H. Akbari, S. Karaman, S. Bhargava, B. Chen, C. Vondrick, and S.-F. Chang, "Multilevel multimodal common semantic space for image-phrase grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12476–12486.

[55] B. Zhang, H. Hu, and F. Sha, "Cross-modal and hierarchical modeling of video and text," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 374–390.

[56] S. Mai, Y. Sun, and H. Hu, "Curriculum learning meets weakly supervised multimodal correlation learning," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2022, pp. 3191–3203.

[57] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 766–775, Apr.–Jun. 2018.

[58] Z. He, L. Wang, and Z. Cai, "Clustered federated learning with adaptive local differential privacy on heterogeneous IoT data," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 137–146, Jan. 2024.

[59] H. Xu, Z. Cai, and W. Li, "Privacy-preserving mechanisms for multilabel image recognition," *ACM Trans. Knowl. Disc. Data*, vol. 16, no. 4, pp. 1–21, 2022.

[60] X. Zheng, L. Zhang, K. Li, and X. Zeng, "Efficient publication of distributed and overlapping graph data under differential privacy," *Tsinghua Sci. Technol.*, vol. 27, no. 2, pp. 235–243, Apr. 2022.

[61] W. Zhang, Z. Xie, A. M. V. V. Sai, Q. Zia, Z. He, and G. Yin, "A local differential privacy trajectory protection method based on temporal and spatial restrictions for staying detection," *Tsinghua Sci. Technol.*, vol. 29, no. 2, pp. 617–633, Apr. 2024.

[62] X. Zheng and Z. Cai, "Privacy-preserved data sharing toward multiple parties in industrial IoTs," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 968–979, May 2020.

[63] D. Olson, "From utterance to text: The bias of language in speech and writing," *Harvard Educ. Rev.*, vol. 47, no. 3, pp. 257–281, 1977.

[64] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, Jul. 1997.

[65] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," 2016, *arXiv:1608.06019*.

[66] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2058–2065.

[67] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptogr. Techn.*, 2006, pp. 486–503.

[68] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate laplace distribution," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 300–303, May 2006.

[69] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnberable: Differential privacy under dependent tuples," in *Proc. NDSS*, 2016, pp. 21–24.

[70] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.

[71] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.

[72] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," 2021, *arXiv:2109.00412*.

[73] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2019, pp. 6558–6569.

[74] E. L. Rosenberg and P. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford, U.K.: Oxford Univ. Press, 2020.

[75] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP-a collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.

[76] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

Honghui Xu (Member, IEEE) received the bachelor's degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2019, and the Ph.D. degree in computer science from Georgia State University, Atlanta, GA, USA, in 2023.

He is currently an Assistant Professor of Information Technology with Kennesaw State University, Kennesaw, GA, USA. His research focuses on trustworthy AI, communication-efficient AI, and Internet of Things.

Dr. Xu was a recipient of the Best Paper Award of the IEEE SmartData 2022 and the Outstanding Research Award from Georgia State University.

Wei Li (Member, IEEE) received the Ph.D. degree in computer science from George Washington University, Washington, DC, USA, in 2016, and the M.S. degree in computer science from Beijing University of Posts and Telecommunications, Beijing, China, in 2011.

She is currently an Associate Professor with the Department of Computer Science, Georgia State University, Atlanta, GA, USA. Her current research spans the areas of blockchain technology, security and privacy for Internet of Things and cyber–physical systems, secure and privacy-aware computing, big data, game theory, and algorithm design and analysis.

Dr. Li won the Best Paper Awards in ACM MobiCom Workshop CRAB 2013 and International Conference WASA 2011. She is a member of ACM.

Daniel Takabi (Member, IEEE) received the B.S. degree in computer engineering from the Amirkabir University of Technology, in 2004, the M.S. degree in information technology from the Sharif University of Technology in 2007, and the Ph.D. degree in information science and technology from the University of Pittsburgh in 2013.

He was an Associate Professor of Computer Science and the Next Generation Scholar with Georgia State University, Atlanta, GA, USA, and was also a Founding Director of the Information Security and Privacy: Interdisciplinary Research and Education Center which is designated as the National Center of Academic Excellence in Cyber Defense Research. He is currently a Professor and the Batten Endowed Chair with the School of Cybersecurity, Old Dominion University, Norfolk, VA, USA. His research interests include various aspects of cybersecurity and privacy, including trustworthy AI, Information security and privacy, and usable security and privacy.

Prof. Takabi is a member of ACM.

Daehee Seo (Member, IEEE) received the B.S. degree in electronic and electrical engineering from Dongshin University, Naju, South Korea, in February 2001, and the M.S. degree in computer science and engineering and the Ph.D. degree in computer science from Soonchunhyang University, Asan, South Korea, in February 2003 and February 2006, respectively.

He currently serves as an Associate Professor with the Faculty of College of Intelligence Information Engineering, Sangmyung University, Seoul, South Korea. His research interests focus on cryptography, AI, bigdata, and blockchain.

Zhipeng Cai (Fellow, IEEE) received B.S. degree from Beijing Institute of Technology, Beijing, China, in 2001, and the M.S. and Ph.D. degrees from the Department of Computing Science, University of Alberta, Edmonton, AB, Canada, in 2004 and 2008, respectively.

He is currently a Professor with the Department of Computer Science, Georgia State University (GSU), Atlanta, GA, USA. Prior to joining GSU, he was a Research Faculty with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. His research areas focus on Internet of Things, machine learning, cyber-security, privacy, networking and big data.

Dr. Cai is the recipient of an NSF CAREER Award. He served as a Steering Committee Co-Chair and a Steering Committee Member for WASA and IPCCC. He also served as a Technical Program Committee Member for more than 20 conferences, including INFOCOM, ICDE, and ICDCS. He has been serving as an Associate Editor-in-Chief for *High-Confidence Computing Journal* (Elsevier) and an Associate Editor for more than ten international journals, including IEEE Internet of Things Journal, IEEE Transactions on Knowledge and Data Engineering, and IEEE Transactions on Vehicular Technology.