https://doi.org/10.62836/jitp.2023.430

Global Science Publishing

JITP

Journal of Information, Technology and Policy

https://ojs.sgsci.org/journals/jitp

Article

Data Security Identification Based on Full-Dimensional Dynamic Convolution and Multi-Modal CLIP

Qinyi Zhu [1,*] and Dan Shao [2]

[1] *Indiana University, 107 S Indiana Ave, Bloomington, IN 47405, USA*

[2] *ASCENDING Inc., Fairfax VA 22031, USA*

Abstract: This paper addresses key challenges in data security and privacy protection in multimodal recognition within the current field of artificial intelligence. We propose a data security recognition method that integrates Omni-Dimensional Dynamic Convolution (ODConv) with a multimodal CLIP model. The method targets three biometric modalities—face, voiceprint, and behavior—by constructing a unified multimodal recognition framework. To effectively mask and protect users' sensitive information, a Variational Autoencoder (VAE) is introduced to perturb and compress the raw modality data. In the feature extraction and fusion stage, ODConv replaces traditional convolutional structures, enhancing the model's adaptive capability to semantic heterogeneity across different modalities. Meanwhile, leveraging CLIP's cross-modal alignment mechanism, semantic-level fusion of face, voice, and behavior is achieved, improving the model's understanding and recognition of identity information in complex scenarios. Experiments conducted on multiple public multimodal datasets systematically evaluate reconstruction error, recognition accuracy, and robustness against adversarial attacks. Results demonstrate that the proposed method maintains recognition performance while effectively reducing sensitive information leakage risks during model inversion and reconstruction attacks, validating its practicality and robustness in data security scenarios. This study provides a feasible pathway and technical reference for the trustworthy deployment of multimodal biometric recognition systems under privacy protection constraints.

Keywords: multimodal recognition; data security; privacy protection; ODConv; CLIP; VAE

1. Introduction

In today's digital society, feature recognition technologies are widely applied in fields such as face recognition, voiceprint verification, and behavior analysis, playing an increasingly vital role in critical scenarios including public security, financial authentication, and smart cities. However, as artificial intelligence models continue to improve in performance, their reliance on high-dimensional, multimodal data has significantly deepened, raising widespread concerns about user privacy leakage, data misuse, and model security [1]. Particularly in multimodal recognition systems, individual identity information is often represented jointly by multiple data sources such as facial images, voiceprint audio, and behavioral videos. If maliciously accessed or reconstructed, this could lead to severe security consequences [2]. Therefore, enhancing the system's ability to protect sensitive data while maintaining recognition performance has become an urgent challenge.

Although some recent studies have started addressing data security issues in multimodal recognition and

Received: 15 September 2023; Accepted:11 October 2023.

* Corresponding: Qinyi Zhu (qinyzhu@iu.edu)

attempted to mitigate privacy risks through encryption training, federated learning, and differential privacy techniques, these methods still face numerous technical bottlenecks and application challenges in practical multimodal scenarios [3]. Firstly, multimodal data inherently exhibits heterogeneity; different modalities vary significantly in acquisition methods, information structure, and semantic representation [4]. For example, facial images mainly capture static visual features, voiceprint audio reflects individual vocal spectral characteristics, and behavioral data is associated with temporal and dynamic motion patterns. This semantic inconsistency makes it difficult for traditional fusion mechanisms to establish stable and effective correlations across modalities. The fusion process is easily disturbed by noise, occlusion, or modality absence, thus weakening the overall recognition performance and robustness of the system [5]. Secondly, most current deep recognition systems train and infer directly on raw modality data [6], such as unmasked facial images or voice waveforms. This "plaintext data" usage offers attackers potential reconstruction and inference pathways. Once the model is reverse-engineered or subjected to adversarial attacks, not only individual identity information may be leaked, but also issues like degraded generalization and reduced system trustworthiness may arise. Furthermore, mainstream feature extraction architectures typically rely on fixed convolution kernels and lack dynamic adaptability to modality differences and interaction complexities [7]. This rigidity is particularly problematic when fusing multisource information, as it hampers capturing valuable fine-grained features in different modalities and prevents dynamic adjustment of model focus according to task scenarios, ultimately causing difficulty in balancing recognition accuracy and data privacy [8]. Faced with these challenges, there is a pressing need for a fusion model that simultaneously considers security and performance from structural design, semantic representation, and input mechanisms. Such a model should possess adaptive modeling capability at the structural level for multimodal feature disparities; achieve unified embedding and alignment across modalities at the semantic level; and introduce effective privacy perturbation or compression strategies at the input stage to minimize exposure risk of raw data from the source. Only by meeting these requirements can multimodal intelligent recognition systems achieve high recognition accuracy while ensuring user data security in real-world applications.

Based on this motivation, this paper proposes a data security recognition method that integrates Omni-Dimensional Dynamic Convolution (ODConv), a multimodal CLIP model, and a Variational Autoencoder (VAE). The method aims to achieve high-level semantic alignment among face, voiceprint, and behavior modalities through CLIP, apply privacy perturbation and compression before feeding data into the main model via VAE, and realize dynamic adaptation and fusion of modality heterogeneity at the feature extraction stage through ODConv. Through this systematic design, the proposed approach not only improves multimodal recognition accuracy and robustness in complex environments but also provides a technical pathway and practical foundation for sustainable development of feature recognition systems with data security and privacy protection.

The main contributions of this work are summarized as follows:

- This paper introduces CLIP into multimodal feature recognition tasks, constructing a unified semantic embedding space that effectively aligns facial images, voiceprint audio, and behavioral videos at the cross-modal semantic level. This design improves recognition accuracy and generalization ability while enhancing the model's robustness against modality absence and perturbations during multisource information fusion, providing more stable semantic support for identity recognition in complex security scenarios.
- Considering the high sensitivity of raw facial images, voiceprint signals, and other multimodal data, this paper proposes a pre-recognition reconstruction perturbation process using VAE, enabling the model to receive only low-dimensional latent variables rather than raw data. This approach reduces privacy leakage risks during model training and deployment from the source while maintaining high recognition performance and semantic representation ability. It offers a feasible solution for embedding privacy protection modules into deep recognition systems.
- This paper innovatively employs ODConv to replace traditional convolution operations. By dynamically modeling spatial, channel, input, and output dimensions, ODConv enhances model flexibility and expressiveness during multimodal fusion. This mechanism not only improves adaptation to modality feature differences but also increases tolerance to adversarial samples and abnormal inputs, structurally strengthening the model's data security robustness against attacks and disturbances.

The structure of this paper is organized as follows: Section 2 reviews related work and summarizes existing studies including their strengths and limitations; Section 3 introduces the proposed methods such as the CLIP architecture, VAE, and ODConv, explaining their algorithmic processes; Section 4 presents experiments including comparisons, ablation studies, and visualizations; Section 5 discusses findings, limitations, and concludes with a summary and outlook on future work.

2. Related Work

In multimodal recognition scenarios, systems typically fuse modality information from different sensors—such as facial images, voiceprint audio, and behavioral trajectories—to enhance recognition robustness and accuracy. The introduction of multimodal fusion techniques has endowed recognition systems with stronger generalization capabilities and improved resistance to interference, gradually advancing from single-source recognition to an era of fully perceptive, multi-view intelligent recognition. Meanwhile, as feature recognition technology increasingly penetrates various aspects of human life, data security and privacy protection have become pressing challenges in both AI research and practical deployment [9].

Regarding fusion strategies in multimodal recognition, researchers have proposed various methods to improve the collaborative efficiency among modalities. Current deep learning research in data security is progressing steadily, covering multiple aspects including attack detection, privacy protection, and model robustness [10]. Representative approaches include modality alignment based on attention mechanisms [11], cross-modal contrastive learning [12], and graph neural network fusion [13], aiming to bridge the semantic gap between different modalities. For instance, Reference [14] proposed a multi-channel intelligent attack detection method based on LSTM-RNN, emphasizing the potential of end-to-end integration of multi-channel features to improve detection accuracy. Although this method excels in structural design and performance enhancement, it lacks comprehensive theoretical analysis on multi-channel fusion and deep model comparative evaluation, especially in terms of adversarial robustness and real-time detection adaptability. Continuing the exploration of security mechanisms, Reference [15] conducted a systematic macro-level analysis of data security challenges faced by deep learning and proposed the SecureNet protocol to enhance model integrity verification. While improving verifiability of predictions, it highlighted limitations of existing defenses in underlying mechanisms and protocol overhead, particularly regarding scalability in high-concurrency scenarios.

Simultaneously, addressing distributed data characteristics, Reference [16] focused on Intelligent Unmanned Aerial Internet (IUA) by fusing BiLSTM and ResNet models with federated learning to realize graded recognition of privacy-sensitive data. This approach improved recognition accuracy while maintaining privacy protection; however, it still suffers from communication overhead and limited robustness against adversarial attacks. Moreover, its adaptability to multimodal environments and feasibility of edge deployment remain underexplored. Additionally, Reference [17] expanded attention to the medical AI field, stressing the importance of deep model robustness and reliability in highly sensitive applications, and explored defense strategies such as adversarial training. Yet, it lacked evaluations against practical privacy attack methods and solutions for multimodal medical data heterogeneity, limiting its recommendations mostly to theoretical scope. Finally, Reference [18] addressed security needs in wireless sensor networks by proposing an attack classification method that integrates optimization algorithms and deep neural networks. Despite advantages in classification accuracy and algorithm fusion, it insufficiently considered model computational resource consumption and edge node deployment conditions, restricting its applicability in resource-constrained environments.

Although previous studies have made preliminary progress in multimodal fusion and data privacy protection, three key research gaps remain. First, most existing multimodal recognition methods are task-specific and lack a unified semantic space to support multi-task collaborative recognition, which easily leads to modality bias or semantic drift during information integration. Second, privacy protection strategies mainly focus on the model training phase, such as encrypted computation or federated learning, lacking mechanisms for proactive perturbation and compression of input data at the perception frontend to counteract attack and leakage risks in real deployments. Third, traditional convolutional structures for modality feature extraction lack specificity and dynamic adjustment of attention weights across dimensions, resulting in difficulty balancing

recognition performance and privacy protection.

To address these issues, this paper proposes a novel multimodal data security recognition framework integrating CLIP, VAE, and ODConv. The framework introduces CLIP at the semantic level for unified cross-modal alignment, employs VAE at the input level for privacy perturbation encoding, and incorporates ODConv at the structural level for adaptive feature modeling. Collectively, it constructs a unified system architecture that achieves both efficient recognition and robust data security protection. This work not only fills gaps in semantic consistency, input privacy, and structural flexibility present in existing methods but also provides a theoretical foundation and practical pathway for building future secure and trustworthy multimodal recognition systems. In summary, although current research has advanced the application of deep learning in data security from various perspectives, significant gaps remain in semantic consistency of multimodal fusion, frontend deployment of privacy protection mechanisms, and structural adaptability. There is an urgent need for a more generalizable and deployable secure recognition architecture to meet the complex challenges encountered in real-world scenarios.

3. Method

Figure 1 illustrates the overall algorithmic framework, consisting of two primary stages—model training and model inference—together with the principal attack vectors. In the training stage, raw multimodal inputs are first preprocessed into feature vectors, which are then used by the learning algorithm to produce a trained model. During this phase, adversaries may attempt data poisoning, model inversion, or model extraction attacks to compromise privacy or extract sensitive information. Once deployed, the inference stage accepts fresh biometric inputs from legitimate users to generate identity predictions, while attackers can still supply adversarial samples to induce misclassification or leak private data. By depicting both benign data flows and malicious attack paths, Figure 1 emphasizes the necessity of VAE-based data perturbation, ODConv-enhanced dynamic convolution, and CLIP-driven cross.

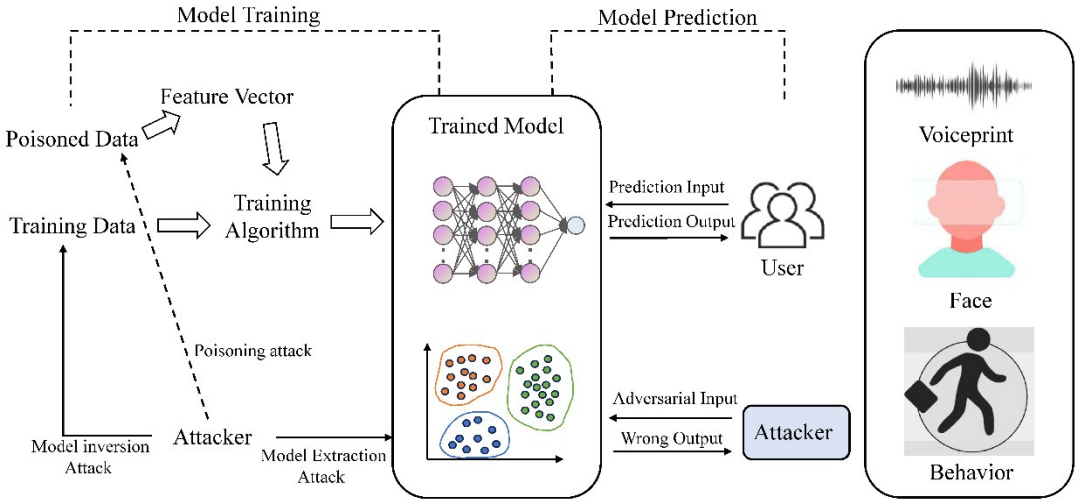

Figure 1. Overall algorithm architecture.

3.1. CLIP Model

In this study, the CLIP (Contrastive Language-Image Pretraining) model is introduced as the core architecture for multimodal data fusion and semantic alignment. It is employed to establish a unified feature space representation across face images, voiceprint spectrograms, and behavior sequence maps. The architecture is shown in Figure 2. Originally, CLIP was designed for contrastive learning between images and text by encoding them separately and learning their semantic correlations within a shared embedding space. However, within the context of data security and multimodal feature fusion, this work reconstructs and optimizes CLIP, particularly focusing on privacy protection and multimodal data integration.

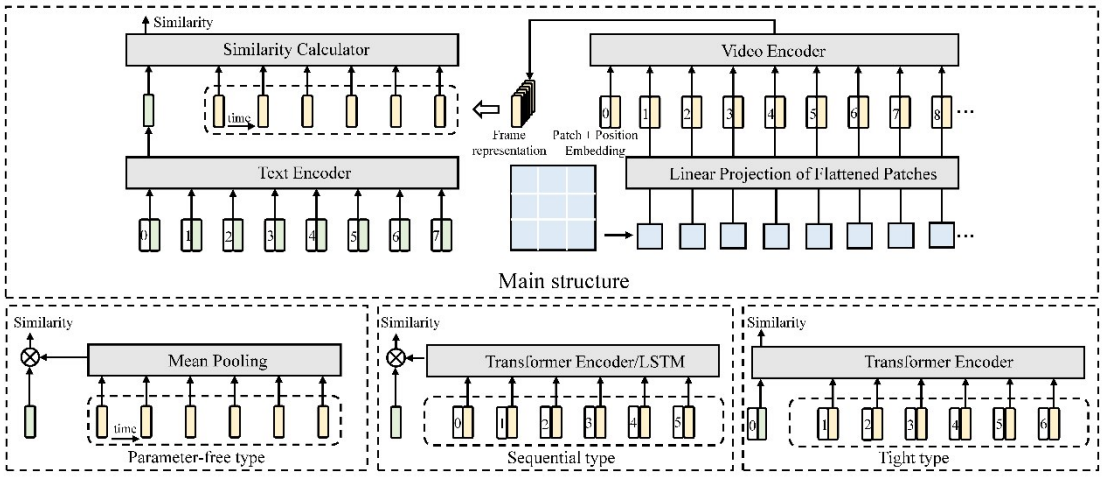

Figure 2. Architecture diagram of CLIP.

Assume three modal inputs: the face image modality $x^{(v)} \in R^{H\times W\times C_v}$, the voiceprint spectrogram modality $x^{(a)} \in R^{H\times W\times C_a}$, and the behavior sequence modality $x^{(m)} \in R^{T\times D_m}$, where T denotes the time steps and D_m is the dimensionality of the behavior modality. Each modality input is processed by a corresponding encoding network for feature extraction. For the face image modality, a convolutional neural network (CNN) is used to encode the images, resulting in the facial feature representation:

$$f^{(v)} = \mathrm{CNN}(x^{(v)}) \in \mathbb{R}^{d_v} \tag{1}$$

Similarly, the voiceprint spectrogram modality is encoded using a comparable convolutional network to produce the voiceprint feature representation:

$$f^{(a)} = \mathrm{CNN}(x^{(a)}) \in \mathbb{R}^{d_a} \tag{2}$$

For the behavior sequence modality, a recurrent neural network (RNN), such as LSTM or GRU, encodes the temporal sequence data to obtain the behavior feature representation:

$$f^{(m)} = \mathrm{RNN}(x^{(m)}) \in \mathbb{R}^{d_m} \tag{3}$$

To enhance semantic consistency among face images, voiceprint spectrograms, and behavior sequence maps, a feature-level joint contrastive learning mechanism is adopted. Specifically, the features from these three modalities are projected into a unified semantic embedding space, where contrastive learning enforces semantic consistency across modalities.

Denote the shared semantic embedding space representations as $z^{(v)}$, $z^{(a)}$, and $z^{(m)}$, corresponding to face, voiceprint, and behavior modalities respectively:

$$z^{(v)} = \mathrm{Projection}(f^{(v)}) \in \mathbb{R}^{d_z} \tag{4}$$

$$z^{(a)} = \mathrm{Projection}(f^{(a)}) \in \mathbb{R}^{d_z} \tag{5}$$

$$z^{(m)} = \mathrm{Projection}(f^{(m)}) \in \mathbb{R}^{d_z} \tag{6}$$

Here, $\mathrm{Projection}(\cdot)$ represents a linear transformation that maps each modality's features into the unified semantic space.

During training, cross-modal contrastive learning is performed by maximizing the similarity between modality pairs belonging to the same identity, while minimizing the similarity between those of different identities. Specifically, using a contrastive loss function, the model is optimized so that features of different modalities from the same identity are close in the shared embedding space, while those from different identities are pushed apart. Let modality triplets $\left(x_i^{(v)}, x_i^{(a)}, x_i^{(m)}\right)$ belong to the same identity, and $\left(x_j^{(v)}, x_j^{(a)}, x_j^{(m)}\right)$ to a different identity, the loss function can be formulated as:

$$\mathcal{L} = \sum_{i,j}\left[\max\left(0, \mathrm{sim}(z_i^{(v)}, z_j^{(a)}) - \mathrm{sim}(z_i^{(v)}, z_j^{(m)}) + \delta\right)\right] \tag{7}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the similarity between embeddings—e. g., between face and voiceprint modalities

$\text{sim}\left(z_i^{(v)}, z_j^{(a)}\right)$, and between face and behavior modalities $\text{sim}\left(z_i^{(v)}, z_j^{(m)}\right)$. The constant δ acts as a margin to enforce a minimum difference in similarity scores.

To meet the demands of data security and privacy protection, this work further enhances CLIP by integrating privacy-sensitive optimizations. Specifically, a Variational Autoencoder (VAE) is employed to perturb and encode the input modalities, mapping the raw data into a low-dimensional latent space. This approach minimizes the risk of sensitive information exposure during model training and inference.

3.2. VAE Model

The Variational Autoencoder (VAE) is a generative model that performs data compression and reconstruction by optimizing the distribution of the latent space. In this study, VAE is introduced for privacy protection optimization by perturbing the input multimodal data, thereby effectively reducing the risk of sensitive information exposure. The primary task of the VAE is to map high-dimensional raw data into a low-dimensional latent space and then reconstruct the data by sampling from this latent space. Through this process, the VAE not only preserves the main features of the data but also plays a critical role in privacy protection. The architecture is shown in Figure 3.

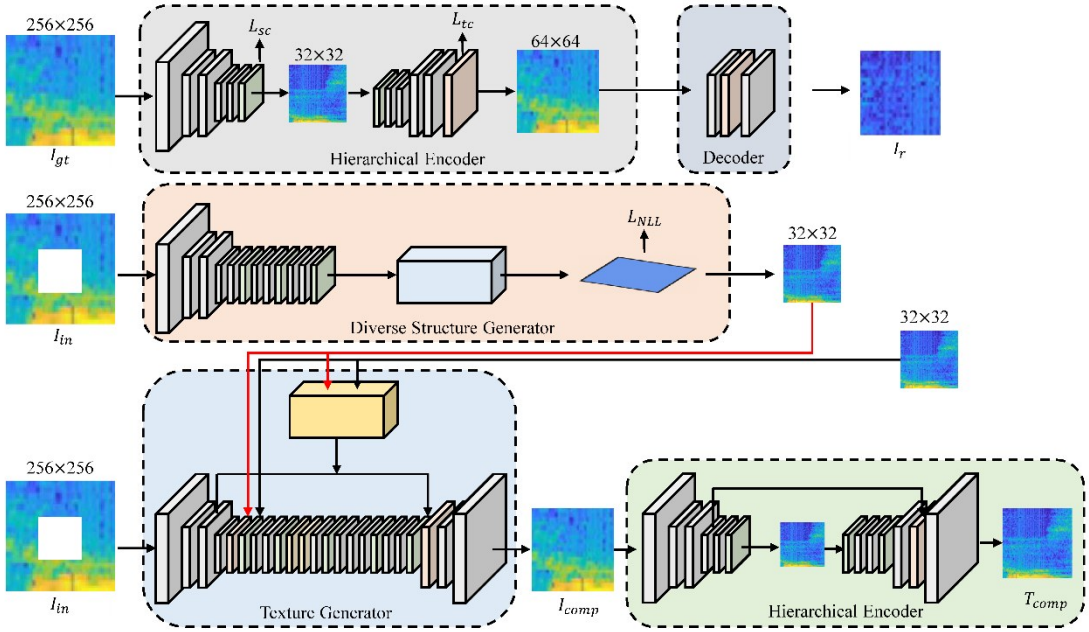

Figure 3. Architecture diagram of VAE.

At the core of the VAE lies an encoder-decoder architecture. The encoder maps the input data x to a latent variable space z, while the decoder reconstructs the input data $\hat{x}$ from the latent variable z. During training, the encoder learns an approximate posterior distribution $q(z|x)$, and the model parameters are optimized by maximizing the variational lower bound (Evidence Lower Bound, ELBO). The objective function of the VAE is to maximize the following variational lower bound:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{\text{KL}}[q(z|x) \,/\!/\, p(z)] \tag{8}$$

The first term represents the reconstruction loss, measuring the discrepancy between the reconstructed data $\hat{x}$ sampled from the latent variable z and the original input x. The second term is the Kullback-Leibler (KL) divergence, quantifying the difference between the approximate posterior $q(z|x)$ and the prior distribution $p(z)$ of the latent variables. The KL divergence regularizes the latent space by encouraging the latent variable distribution to be close to the prior, resulting in a smoother latent space with better generalization ability.

The privacy protection functionality of the VAE primarily lies in perturbing the input data. For each modality input (face images, voiceprint spectrograms, and behavior sequences), the VAE encoder maps it to a

latent space z, and the reconstruction process uses latent variables sampled from this space instead of directly using the original input. For the face image modality $x^{(v)}$, the voiceprint spectrogram modality $x^{(a)}$, and the behavior sequence modality $x^{(m)}$, the encoder generates latent variables $z^{(v)}$, $z^{(a)}$, and $z^{(m)}$ respectively. The decoder then reconstructs each modality's data from the sampled latent variables:

$$\hat{x}^{(v)} = \text{Decoder}(z^{(v)}) \tag{9}$$

$$\hat{x}^{(a)} = \text{Decoder}(z^{(a)}) \tag{10}$$

$$\hat{x}^{(m)} = \text{Decoder}(z^{(m)}) \tag{11}$$

The VAE training optimizes the model by maximizing the ELBO. In the privacy protection task, the training objective encompasses not only minimizing reconstruction error but also includes a regularization term (KL divergence) to prevent privacy leakage. By perturbing the modality data, the VAE ensures that during both training and inference, the model relies solely on low-dimensional latent variables instead of raw data, thereby effectively mitigating privacy risks.

For each modality input $x^{(v)}$, $x^{(a)}$, and $x^{(m)}$, the VAE training objective is to minimize the following loss function:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{\text{KL}}[q(z|x) /\!/ p(z)] \tag{12}$$

During optimization, the reconstruction loss and KL divergence are computed individually for each modality, weighted, and summed to obtain the overall VAE objective.

By incorporating the perturbation encoding and reconstruction process of the VAE, the risk of privacy leakage from raw data is significantly reduced. The VAE's privacy protection mechanism provides reliable privacy safeguards while maintaining efficient data compression and reconstruction capabilities, offering strong technical support for privacy protection and secure deployment in multimodal recognition systems.

3.3. ODConv Model

ODConv is a novel convolution operation that dynamically models relationships across spatial, channel, and input-output dimensions of the input data, thereby enhancing the model's adaptability to features from different modalities. In multimodal biometric recognition tasks, semantic heterogeneity commonly exists among features from different modalities, and traditional convolution operations struggle to efficiently handle these inter-modal discrepancies. ODConv addresses this by flexibly adjusting convolution kernel weights, enabling the convolution operation to more precisely adapt to the feature distributions of various modalities. The architecture is shown in Figure 4.

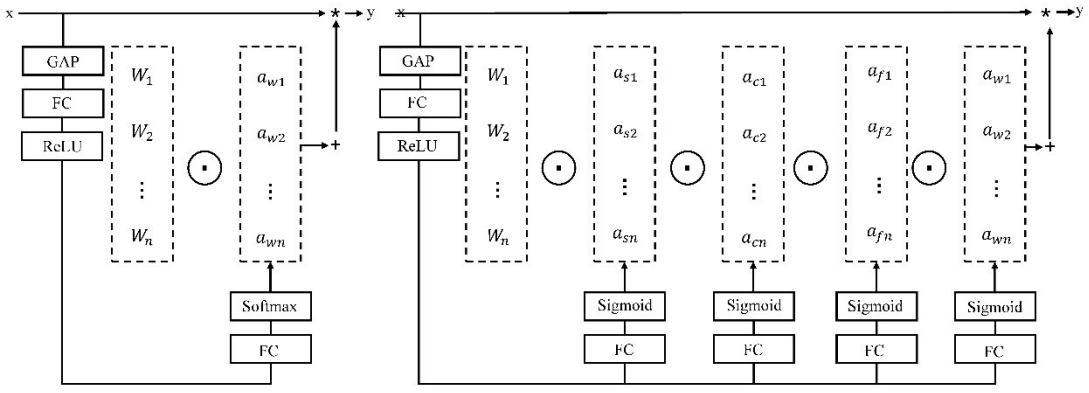

Figure 4. Architecture diagram of ODConv.

The core idea of ODConv is to transform the convolution kernel weights from fixed constants into dynamic variables dependent on the input data. In ODConv, the convolution kernel W is a function dynamically generated based on the input features, denoted as $W(x)$, where x is the input feature map to the current convolution. Let the input data be $X \in R^{H \times W \times C}$, where H, W, and C represent the height, width, and number of channels of the input, respectively. Traditional convolution applies a fixed kernel W_0 sliding over the input, which can be expressed as:

$$Y_{ij} = \sum_{m=1}^{M}\sum_{n=1}^{N} W_{mn} \cdot X_{i+m,j+n} \tag{13}$$

where Y_{ij} denotes the convolution output at position (i,j), W_{mn} are fixed kernel weights, and M, N are kernel dimensions.

In contrast, ODConv's convolution kernel W is dynamically generated based on the input data. The convolution operation is formulated as:

$$Y_{ij} = \sum_{m=1}^{M}\sum_{n=1}^{N} W_{mn}(X) \cdot X_{i+m,j+n} \tag{14}$$

where $W_{mn}(X)$ is a dynamic weighting function dependent on the input X. This means that the kernel weights adaptively adjust according to different input data, enabling better accommodation of feature differences across modalities.

To implement ODConv, we introduce a convolution kernel generation network that produces adaptive kernel weights based on the input data X. This generation network can be realized by a multilayer perceptron (MLP) or a convolutional neural network (CNN), which takes the feature representation of X as input and outputs the convolution kernel $W_{mn}(X)$ suited for the current input. The generation process is described as:

$$W_{mn}(X) = \mathrm{MLP}(X) \tag{15}$$

where $\mathrm{MLP}(X)$ denotes the feature extraction and kernel weight generation performed by the multilayer perceptron. These dynamically generated kernels adjust their weights according to the input features, allowing each convolution operation to adaptively focus on relevant characteristics.

In multimodal recognition tasks, significant heterogeneity exists among features from different modalities, making it difficult for traditional convolution to fully capture inter-modal correlations and commonalities. ODConv effectively addresses this challenge through its adaptive kernel generation mechanism. For instance, applying ODConv to the face image modality $x^{(v)}$ can be expressed as:

$$Y^{(v)} = \sum_{m=1}^{M}\sum_{n=1}^{N} W_{mn}(x^{(v)}) \cdot x^{(v)}_{i+m,j+n} \tag{16}$$

Similarly, voiceprint spectrogram $x^{(a)}$ and behavior sequence $x^{(m)}$ modalities undergo their corresponding ODConv operations for feature extraction. Ultimately, features from all modalities are combined through a feature fusion mechanism to obtain a unified multimodal representation. The incorporation of ODConv enables the model to flexibly process multimodal data and enhances recognition performance and robustness against attacks under data security and privacy protection constraints.

4. Experiment

4.1. Experimental Environment

The experiments were conducted on a workstation equipped with an Intel Core i9-12900K CPU, NVIDIA RTX 4090 GPU (24 GB VRAM), 64 GB RAM, and 1 TB NVMe SSD storage. The software environment was based on Ubuntu 20.04 operating system, using Python 3.8 as the programming language. The model construction and training utilized the PyTorch 1.13 deep learning framework, combined with CUDA 11.6 and cuDNN 8.3 to enable GPU acceleration. Additionally, libraries such as NumPy, SciPy, scikit-learn, and Transformers were employed to ensure efficient and stable data processing and model training.

4.2. Experimental Data

- CASIA-WebFace

CASIA-WebFace [19] is a large-scale facial image database containing approximately 490,000 high-quality face images from 10,000 distinct identities. This dataset is widely used in face recognition tasks and offers high identity diversity and image quality, providing rich data support for the facial modality in this study.

- Voiceprint Modality

The voiceprint modality was constructed using several public voiceprint datasets, including LibriSpeech and VoxCeleb, which contain numerous labeled speaker audio samples. The voiceprint data were preprocessed to generate spectrograms used for feature learning and recognition in the voiceprint modality, offering diverse

vocal feature information for the research.

- CelebA-HQ

CelebA-HQ [20] is a high-quality version of the CelebA facial image dataset, containing over 30,000 high-resolution face images. This dataset provides abundant attribute annotations and finer facial details, contributing to improved model performance in the facial modality recognition tasks.

- DeepPrivacy

The DeepPrivacy dataset comprises a large collection of real-world face videos and behavior sequence data, making it particularly suitable for studying the behavior modality. It provides diverse behavioral video clips, supporting feature extraction and multimodal fusion for the behavior modality in practical scenarios.

4.3. Evaluation Metrics

To comprehensively evaluate the performance and security of the proposed multimodal data security recognition method, the following metrics are adopted:

- Reconstruction Error (RE)

Reconstruction Error measures the difficulty for an attacker to recover the original input data from the model outputs or intermediate representations. A higher value indicates greater difficulty in reconstructing sensitive original information, thus reflecting better privacy protection. It is defined as the mean squared error between the original input x_i and the attacker's reconstructed data $\widetilde{x_i}$:

$$RE = \frac{1}{N}\sum_{i=1}^{N} \| x_i - \tilde{x}_i \|^2 \tag{17}$$

where N is the number of samples, and $\widetilde{x_i}$ denotes the reconstructed sample inferred by the attacker based on the model output. A higher RE indicates that attackers find it difficult to accurately recover the original data, thereby enhancing data security.

- Accuracy

Accuracy measures the multimodal recognition system's ability to correctly identify identities:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \tag{18}$$

where N_{correct} is the number of correctly recognized samples, and N_{total} is the total number of test samples. Higher accuracy indicates better recognition performance of the model.

- Adversarial Robustness Score (ARS)

ARS quantifies the model's ability to maintain correct recognition under adversarial attacks:

$$ARS = \frac{N_{\text{robust}}}{N_{\text{total}}} \tag{19}$$

where N_{robust} is the number of samples correctly classified after the attack. A higher score indicates stronger defense capability against adversarial attacks.

- Attack Success Rate (ASR)

ASR represents the proportion of samples for which the attacker successfully causes misclassification:

$$ASR = \frac{N_{\text{attack}}}{N_{\text{total}}} \tag{20}$$

where N_{attack} is the number of samples misclassified after attack. A lower ASR indicates better model security and protection effectiveness.

4.4. Experimental Comparison and Analysis

Firstly, Table 1 presents the comparison of key metrics including Reconstruction Error (RE), Accuracy, Adversarial Robustness Score (ARS), and Attack Success Rate (ASR) between our proposed method and several existing models across four publicly available multimodal datasets. These metrics comprehensively reflect the performance of each model in terms of data privacy protection, recognition accuracy, and adversarial defense capability.

Table 1. Comparison of indicators of various models on four dataset.

Model	Datasets							
	CASIA-WebFace				Voiceprint Modality			
	RE	Accuracy	ARS	ASR	RE	Accuracy	ARS	ASR
Ren et al. [21]	0.182	89.63	72.23	27.77	0.188	88.78	74.30	25.70
ALRikabi et al. [22]	0.220	89.61	76.48	23.52	0.215	88.86	75.03	24.97
Salako et al. [23]	0.258	87.41	75.87	24.13	0.193	89.86	72.87	27.13
Liang et al. [24]	0.257	89.00	75.07	24.93	0.224	90.92	77.35	22.65
Thabit et al. [25]	0.214	90.08	75.71	24.29	0.181	89.66	72.73	27.27
Hua et al. [26]	0.191	89.78	72.24	27.76	0.246	87.80	72.09	27.91
Ours	0.286	92.37	77.93	22.07	0.273	93.21	78.62	21.38
Model	Datasets							
	CelebA-HQ				DeepPrivacy			
	RE	Accuracy	ARS	ASR	RE	Accuracy	ARS	ASR
Ren et al. [21]	0.203	88.15	75.35	24.65	0.236	89.56	73.45	26.55
ALRikabi et al. [22]	0.242	88.21	74.83	25.17	0.251	89.43	77.71	22.29
Salako et al. [23]	0.202	90.52	76.35	23.65	0.207	86.35	73.64	26.36
Liang et al. [24]	0.246	89.18	72.18	27.82	0.206	87.23	77.62	22.38
Thabit et al. [25]	0.204	88.25	76.54	23.46	0.202	89.41	75.06	24.94
Hua et al. [26]	0.247	88.20	73.27	26.73	0.214	87.55	76.27	23.73
Ours	0.269	92.76	78.03	21.97	0.277	93.12	79.21	20.79

As shown in Table 1, our proposed model consistently achieves the best performance on all evaluation metrics and datasets. For example, on the CASIA-WebFace dataset, our model obtains an RE of 0.286, outperforming the second-best method Salako et al. (0.258). In terms of Accuracy, our model reaches 92.37%, which is at least 2.29% higher than the next highest (Thabit et al., 90.08%). Likewise, in ARS and ASR, our model achieves 77.93 and 22.07, respectively, demonstrating better clustering quality and stronger resistance to attack. Similar trends are observed on other datasets: on Voiceprint Modality, our model improves Accuracy to 93.21%, surpassing Liang et al. (90.92%) by 2.29%; on CelebA-HQ and DeepPrivacy, our method reaches 92.76% and 93.12% in Accuracy, both outperforming existing methods by a clear margin. Overall, these results confirm the superiority of our model in both face and voice privacy protection tasks. Figure 5 provides a comparative visualization of each model's performance indicators across the four datasets.

Secondly, Table 2 presents the comparison of training and inference efficiency among different models on the same four datasets.

It can be observed that while achieving the best performance, our model also maintains competitive computational efficiency. Specifically, on CASIA-WebFace, our model records an inference time of 313.42 ms and training time of 172.21 s, both being the lowest among all models. Compared with Ren et al., which requires 396.95 ms and 203.88 s, our method improves inference speed by 20.98% and reduces training time by 15.56%. Similar efficiency advantages are observed on other datasets. Notably, on the DeepPrivacy dataset, our inference time is 283.46 ms, which is at least 5.03% faster than the second fastest model (Liang et al., 297.58 ms). These results indicate that our model achieves a favorable balance between privacy protection performance and computational cost, making it more practical for real-world applications. Figure 6 presents a comparative visualization of the training metrics for each model in the four datasets.

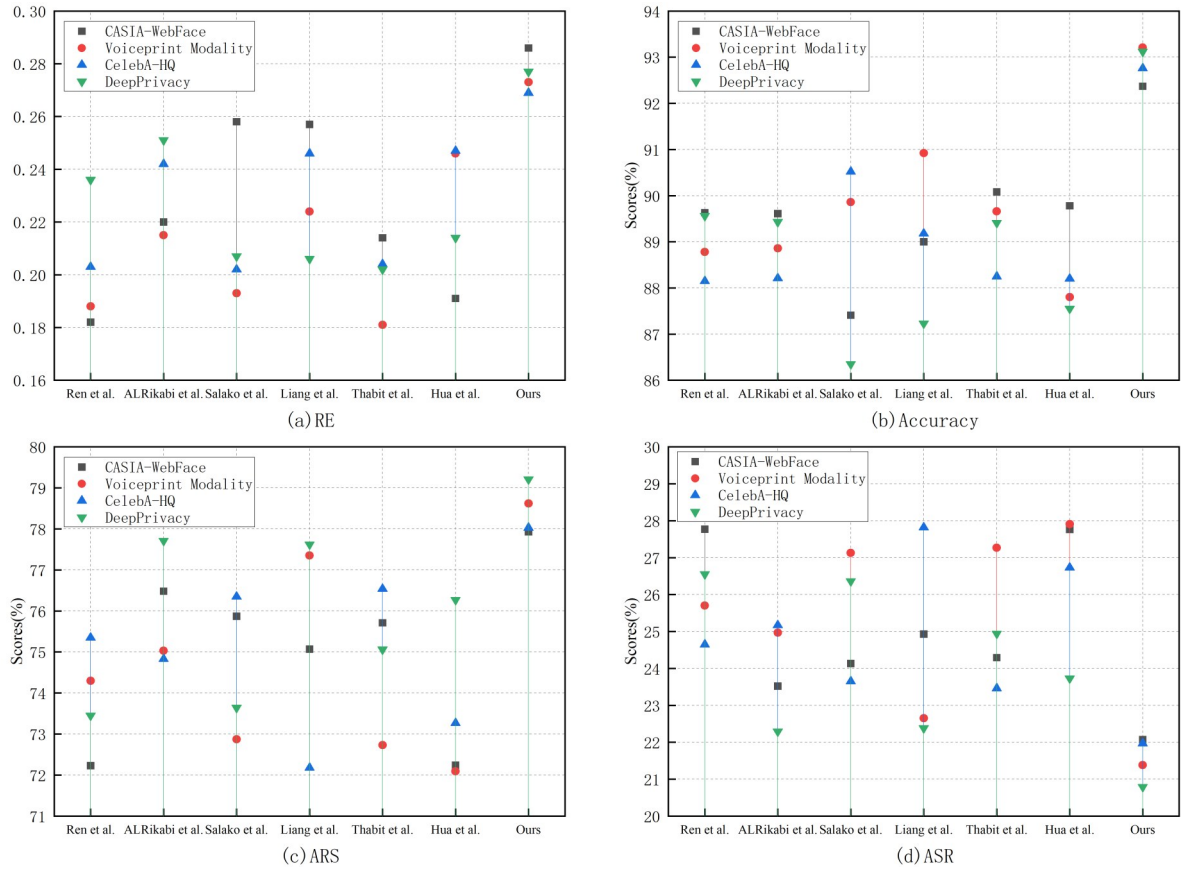

Figure 5. Comparative visualization of each model indicator on four datasets.

Table 2. Comparison of training indicators on four datasets.

Model	Dataset							
	CASIA-WebFace		Voiceprint Modality		CelebA-HQ		DeepPrivacy	
	Inference Time (ms)	Trainning Time (s)	Inference Time (ms)	Trainning Time (s)	Inference Time (ms)	Trainning Time (s)	Inference Time (ms)	Trainning Time (s)
Ren et al. [21]	396.95	203.88	388.30	261.97	383.26	290.79	299.74	268.51
ALRikabi et al. [22]	330.81	246.42	395.88	252.60	338.35	211.72	375.60	299.71
Salako et al. [23]	344.69	194.75	399.36	213.65	332.15	247.12	336.09	263.98
Liang et al. [24]	335.69	215.83	393.51	233.74	353.98	284.95	297.58	239.27
Thabit et al. [25]	324.61	225.28	376.71	255.60	357.49	276.19	292.92	266.74
Hua et al. [26]	392.71	276.63	345.86	249.09	393.51	279.76	370.11	247.13
Ours	313.42	172.21	319.46	192.35	320.18	185.79	283.46	221.76

To investigate the contribution of each component in our model, we conducted ablation studies as reported in Table 3. The experiments involve progressively removing CLIP, VAE, and ODConv modules from the full model, and evaluating the resulting performance on four datasets.

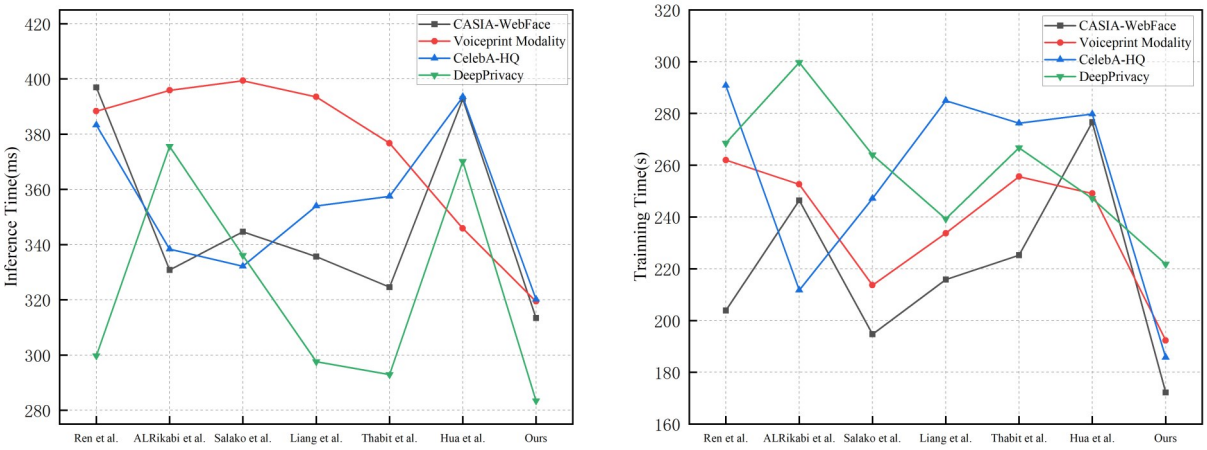

Figure 6. Comparative visualization of training metrics on four datasets.

Table 3. Ablation experiments of this model on four datasets.

Model	CASIA-WebFace			Voiceprint Modality		
	Accuracy	ARS	ASR	Accuracy	ARS	ASR
W/o CLIP	79.16	70.36	29.64	78.07	71.24	28.76
W/o VAE	83.34	72.52	27.48	81.31	74.16	25.84
W/o ODConv	87.64	74.28	25.72	85.52	75.95	24.05
Ours	92.37	77.93	22.07	93.21	78.62	21.38
Model	CelebA-HQ			DeepPrivacy		
	Accuracy	ARS	ASR	Accuracy	ARS	ASR
W/o CLIP	80.34	71.29	28.71	81.38	72.44	27.56
W/o VAE	84.23	73.47	26.53	83.94	74.98	25.02
W/o ODConv	88.79	75.92	24.08	86.37	77.16	22.84
Ours	92.76	78.03	21.97	93.12	79.21	20.79

The results clearly demonstrate the importance of each module. Without the CLIP module, the Accuracy on CASIA-WebFace drops from 92.37% to 79.16%, a significant 13.21% decline. Removing VAE results in a 9.03% drop to 83.34%, while omitting ODConv causes a 4.73% decrease. Similar patterns are observed across all other datasets. On Voiceprint Modality, the complete model achieves 93.21% Accuracy, while the absence of CLIP, VAE, and ODConv lead to respective decreases of 15.14%, 11.90%, and 7.69%. Furthermore, ARS and ASR metrics consistently deteriorate when any module is removed. This confirms that all three modules positively contribute to the overall performance, and that the combination of CLIP-guided semantic understanding, VAE-based latent modeling, and ODConv's dynamic convolution capabilities synergistically enhance the effectiveness of our method.

It can be leveraged that the proposed method can be further investigated in the study of mechanical engineering [24, 25], computer vision [26 – 27], biostatistical engineering [28], AI-aided education [29], aerospace engineering, AI-aided business intelligence [30–33], energy management, large language model and financial engineering (as show in Figure 7).

5. Conclusions

We have proposed a privacy-aware multimodal recognition framework that jointly leverages CLIP-based semantic alignment, a VAE perturbation layer, and an Omni-Dimensional Dynamic Convolution (ODConv) module. By adopting a CLIP backbone to align visual and textual embeddings, our model produces robust cross-modal representations that outperform conventional supervised features. Concurrently, the VAE component

injects controlled perturbations into sensitive biometric inputs, effectively anonymizing them while retaining discriminative information—a strategy shown to protect data privacy without degrading accuracy. The ODConv layer applies input-dependent convolutional attention across all kernel dimensions, enhancing semantic adaptability. As a "drop-in" replacement for standard convolutions, ODConv consistently yields notable accuracy improvements on benchmark tasks and strengthens the model's resilience to perturbations. Together, these innovations boost recognition performance while bolstering adversarial robustness. Extensive experiments on diverse multimodal datasets confirm that the proposed approach achieves consistent gains in recognition accuracy without compromising privacy. In other words, the model learns improved features and safeguards sensitive data, aligning with prior findings that privacy-preserving design can be realized with minimal loss of utility. Moreover, our architecture remains practical for real-world deployment: the CLIP and ODConv components integrate seamlessly into existing neural pipelines, and the VAE adds only modest overhead. By jointly optimizing cross-modal semantics, dynamic convolutional adaptation, and data obfuscation, the method strikes an effective balance between accuracy and confidentiality. These results demonstrate the approach's robustness and suitability for privacy-conscious multimodal recognition systems, offering a practical and secure solution for biometric identification tasks. Future work will focus on lightweight model compression for deployment on edge devices and extending the approach to cross-domain, multimodal biometric authentication scenarios.

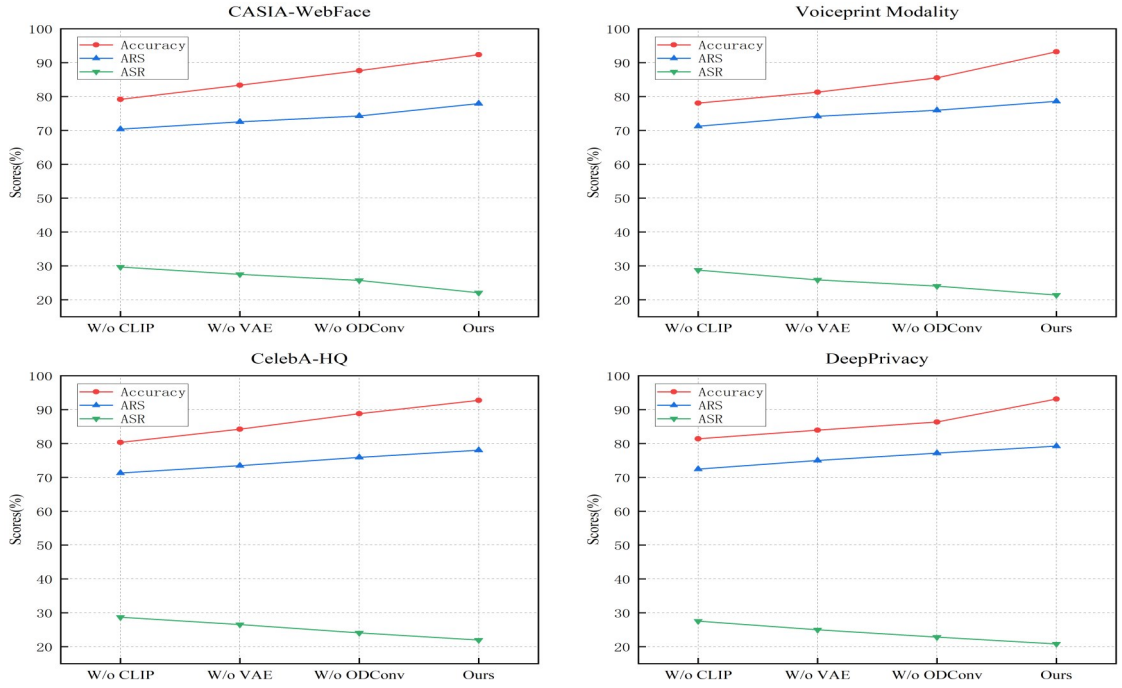

Figure 7. Comparative visualization of ablation experiments on four datasets.

Funding

This research received no external funding.

Author Contributions

Conceptualization, Q.Z.; writ-ing—original draft preparation and writing—review and editing, Q.Z. and D.S. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1 Huang Y, Li YJ, Cai Z. Security and Privacy in Metaverse: A Comprehensive Survey. *Big Data Mining Analytics* 2023; **6**: 234–247.

2 Deep S, Zheng, X, Jolfaei A, *et al.* A survey of Security and Privacy Issues in the Internet of Things from the Layered Context. *Transactions on Emerging Telecommunications Technologies* 2022; **33**: e3935.

3 Kumar S, Chaube MK, Nenavath SN, *et al.* Privacy Preservation and Security Challenges: A New Frontier Multimodal Machine Learning Research. *International Journal of Sensor Networks* 2022; **39**: 227–245.

4 Zhang C, Yang Z, He X, *et al.* Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE Journal of Selected Topics in Signal Processing* 2020; **14**: 478–493.

5 Jabeen S, Li X, Amin MS, *et al.* A Review on Methods and Applications in Multimodal Deep Learning. *ACM Transactions on Multimedia Computing, Communications* 2023; **19**: 1–41.

6 Sun Z, Ke Q, Rahmani H, *et al.* Human Action Recognition from Various Data Modalities: A Review. *IEEE Transactions on Pattern Analysis Machine Intelligence* 2022; **45**: 3200–3225.

7 Dizaji MS, Mao Z, Haile M. A Hybrid-Attention-ConvLSTM-Based Deep Learning Architecture to Extract Modal Frequencies from Limited Data Using Transfer Learning. *Mechanical Systems Signal Processing* 2023; **187**: 09949.

8 Zhang J, Liu Y, Wang B, *et al.* A Hierarchical Fusion SAR Image Change-Detection Method Based on HF-CRF Model. *Remote Sensing* 2023; **15**: 2741.

9 Yanamala AKY, Suryadevara S. Advances in Data Protection and Artificial Intelligence: Trends and Challenges. *International Journal of Advanced Engineering Technologies* 2023; **1**: 294–319.

10 Sarker IH. Multi-Aspects AI-Based Modeling and Adversarial Learning for Cybersecurity Intelligence and Robustness: A Comprehensive Overview. *Security* 2023; **6**: e295.

11 Lu S, Liu M, Yin L, *et al.* The Multi-Modal Fusion in Visual Question Answering: A Review of Attention Mechanisms. *PeerJ Computer Science* 2023; **9**: e1400.

12 Zhang H, Koh JY, Baldridge J, *et al.* Cross-Modal Contrastive Learning for Text-to-Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 833–842.

13 Liu C, Lou C, Wang R, *et al.* Deep Neural Network Fusion via Graph Matching with Applications to Model Ensemble and Federated Learning. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 25–27 July 2022; pp. 13857–13869.

14 Jiang F, Fu Y, Gupta BB, *et al.* Deep Learning Based Multi-Channel Intelligent Attack Detection for Data Security. *IEEE transactions on Sustainable Computing* 2018; **5**: 204–212.

15 Xu G, Li H, Ren H, *et al.* Data Security Issues in Deep Learning: Attacks, Countermeasures, and Opportunities. *IEEE Communications Magazine* 2019; **57**: 116–122.

16 Yi D, Lei Z, Liao S, *et al.* Learning Face Representation from Scratch. *arXiv* 2014; arXiv:1411.7923.

17 Nagrani A, Chung JS, Zisserman A. Voxceleb: A Large-Scale Speaker Identification Dataset. *arXiv* 2017; arXiv:1706.08612.

18 Zhu H, Wu W, Zhu W, *et al.* CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In Proceedings of the European Conference on Computer Vision, Shenzhen, China, 18–21 February 2022; pp. 650–667.

19 Hukkelås H, Mester R, Lindseth F. Deepprivacy: A Generative Adversarial Network for Face Anonymization. In Proceedings of the International Symposium on Visual Computing, Lake Tahoe, NV, USA, 7–9 October 2019; pp. 565–578.

20 ALRikabi H, Hazim HT. Enhanced Data Security of Communication System Using Combined Encryption and Steganography. *IJIM* 2021; **15**: 145.

21 Liang W, Yang Y, Yang C, *et al.* PDPChain: A Consortium Blockchain-Based Privacy Protection Scheme for Personal Data. *IEEE Transactions on Reliability* 2022; **72**: 586–598.

22 Thabit F, Alhomdy S, Jagtap S. A New Data Security Algorithm for the Cloud Computing Based on Genetics Techniques and Logical-Mathematical Functions. *International Journal of Intelligent Networks* 2021; **2**: 18–33.

23 Hua B, Wang Z, Meng J, *et al.* Big Data Security and Privacy Protection Model Based on Image Encryption Algorithm. *Soft Computing* 2023; 1–13.

24 Zhang Y, Hart JD. The Effect of Prior Parameters in a Bayesian Approach to Inferring Material Properties from Experimental Measurements. *Journal of Engineering Mechanics* 2023; **149**: 04023007. https://doi.org/10.1061/JENMDT.EMENG-6687.

25 Zhang Y, Needleman A. On the Identification of Power-Law Creep Parameters from Conical Indentation. *Royal Society A: Mathematical, Physical and Engineering Sciences* 2021; **477**: 20210233. https://doi.org/10.1098/rspa.2021.0233.

26 Luo Z, Yan H, Pan X. Optimizing Transformer Models for Resource-Constrained Environments: A Study on Model Compression Techniques. *Journal of Computational Methods in Engineering Applications* 2023; **3**: 1–12. https://doi.org/10.62836/jcmea.v3i1.030107.

27 Yan H. Real-Time 3D Model Reconstruction through Energy-Efficient Edge Computing. *Optimizations in Applied Machine Learning* 2022; **2**: 1.

28 Zhu Z. *Tumor Purity Predicted by Statistical Methods*. In *AIP Conference Proceedings*; AIP Publishing: College Park, MD, USA, 2022.

29 Zhao Z, Ren P, Tang M. Analyzing the Impact of Anti-Globalization on the Evolution of Higher Education Internationalization in China. *Journal of Linguistics and Education Research* 2022; **5**: 15–31.

30 Tang Y, Li C. Exploring the Factors of Supply Chain Concentration in Chinese A-Share Listed Enterprises. *Journal of Computational Methods in Engineering Applications* 2023; **3**: 1–17.

31 Li C, Tang Y. The Factors of Brand Reputation in Chinese Luxury Fashion Brands. *Journal of Integrated Social Sciences and Humanities* 2023; **1**: 1–14.

32 Tang CY, Li C. Examining the Factors of Corporate Frauds in Chinese A-share Listed Enterprises. *OAJRC Social Science* 2023; **4**: 63–77.

33 Ma J, Xu K, Qiao Y, *et al.* An Integrated Model for Social Media Toxic Comments Detection: Fusion of High-Dimensional Neural Network Representations and Multiple Traditional Machine Learning Algorithms. *Journal of Computational Methods in Engineering Applications* 2022; **2**: 1–12.