PRIVFUSIONNET: MITIGATING IDENTITY RE-LEAKAGE IN EEG–EYE MOVEMENT FUSION FOR EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fusion leverages complementary signals but can also recombine residual subject-specific cues, causing identity information suppressed within individual modalities to re-emerge after fusion. Existing privacy-preserving approaches primarily operate at the unimodal level, leaving fusion-level privacy underexplored. Using emotion recognition from EEG and eye-movement signals as a testbed, we formulate fusion-induced identity re-leakage and show that unimodal identity suppression alone does not reliably protect fused representations. We propose PrivFusionNet, a privacy-preserving multimodal representation learning framework that directly protects the fusion space through three complementary mechanisms. Privacy-aware low-rank fusion reweights cross-modal interaction components according to task utility and identity risk; fusion-level task-identity disentanglement attenuates identity-related directions while retaining emotion-discriminative information; and a fused privacy codebook quantizes continuous representations into subject-shared task prototypes to suppress fine-grained identity cues. Under leave-one-session-out privacy evaluation, PrivFusionNet reduces fused-representation user identification accuracy from 25.33% to 7.57% on SEED-IV and from 46.38% to 6.11% on SEED-V. Under leave-one-subject-out emotion-recognition evaluation, it achieves mean accuracies of 84.23% and 88.30%, respectively. These results identify fusion-induced identity re-leakage as a practical privacy risk and demonstrate a favorable privacy-utility trade-off.

1 INTRODUCTION

Multimodal representation learning integrates information from heterogeneous modalities to obtain richer representations than those derived from a single source (Ngiam et al., 2011; Baltrušaitis et al., 2019). In physiological computing, electroencephalography (EEG) and eye movement signals capture user states from complementary perspectives: EEG reflects neural activity, whereas eye movements characterize visual and behavioral responses. Their combination has been widely used for emotion recognition and other user-state inference tasks (Lu et al., 2015; Zheng et al., 2019). Accordingly, most existing multimodal methods focus on improving task performance through more effective cross-modal alignment, interaction modeling, or fusion (Zadeh et al., 2017; Tsai et al., 2019).

However, this task-centered perspective overlooks an important privacy issue. In addition to task-related information, EEG and eye movement signals also contain subject-specific characteristics. EEG representations can encode stable spectral and spatial patterns that can be exploited for user identification (Meng et al., 2023; Wang et al., 2025). Likewise, eye movement patterns such as scanpaths, fixation statistics, and saccadic behavior have also been shown to carry discriminative identity information (Holland & Komogortsev, 2011; 2013; Lohr et al., 2021; Lohr & Komogortsev, 2022). Therefore, even when identity is irrelevant to the downstream task, intermediate representations produced by multimodal models may still reveal information about the users who generated the data.

A natural strategy is to remove identity information from each modality before fusion. Existing privacy-preserving methods largely follow this direction by learning invariant representations or

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

disentangling task-related and sensitive factors (Xie et al., 2017; Madras et al., 2018; Moyer et al., 2018). However, privacy at the unimodal level does not necessarily remain intact after fusion. A fusion module does more than simply concatenate two representations; it further models dependencies and cross-modal interactions between modalities. As a result, residual identity cues that are weak in individual modalities may become highly discriminative when combined. We refer to this phenomenon as *fusion-induced identity re-leakage*. This raises a fundamental question for multimodal privacy: *does unimodal privacy protection remain effective after multimodal fusion?*

We study this problem using EEG-eye multimodal emotion recognition as a testbed. We first construct a fusion-level identity re-leakage verification framework that compares a standard multimodal model with its unimodal privacy-protected counterpart under the same encoders and fusion backbone. Identity probes are then applied to both unimodal and fused representations to measure how much subject information remains before and after fusion (Coavoux et al., 2018; Madras et al., 2018). The results show that separately reducing identity information in EEG and eye movement representations does not reliably reduce identity leakage in the fused representation. In some settings, the user identification accuracy of the fused representation even increases after unimodal identity suppression, indicating that residual subject-specific cues may be recombined through cross-modal interactions.

Based on this observation, we propose PrivFusionNet, a privacy-preserving multimodal representation learning framework that directly targets fusion-level identity leakage. PrivFusionNet introduces privacy protection at three stages of the fusion pipeline. First, a privacy-aware low-rank fusion module scores cross-modal interaction components according to both task utility and identity risk, thereby reducing the contribution of identity-sensitive interactions. Second, a fusion-level task-identity disentanglement module suppresses identity-related subspaces in the fused representation while preserving task-discriminative information. Finally, a fused privacy codebook maps continuous fused representations to discrete task prototypes shared across subjects, limiting fine-grained subject-specific variations that could support identity inference. Through these mechanisms, PrivFusionNet directly protects fused representations during multimodal interaction and in the resulting representation space, rather than relying solely on unimodal privacy protection.

We evaluate PrivFusionNet on the public SEED-IV and SEED-V EEG-eye emotion datasets under cross-session and strict leave-one-subject-out (LOSO) settings. Across both datasets, PrivFusionNet substantially reduces the user identification accuracy of fused representations while maintaining strong emotion recognition performance. These results suggest that fusion-induced identity re-leakage is a distinct privacy failure mode that cannot be inferred solely from unimodal privacy measurements, highlighting the need for privacy mechanisms explicitly designed for fused representations.

Our main contributions are:

- We identify and empirically analyze fusion-induced identity re-leakage, showing that suppressing identity information in individual modalities does not necessarily make the resulting fused representation private.
- We propose PrivFusionNet, which suppresses identity leakage in fused representations through privacy-aware fusion, fusion-level disentanglement, and discrete prototype constraints, while maintaining strong task performance on two public EEG-eye multimodal datasets.

2 Related Work

2.1 Emotion Recognition

Physiological signals have been widely studied for emotion recognition because they provide complementary observations of internal neural activity and external behavioral responses. The DEAP dataset established an early benchmark for emotion recognition from EEG and peripheral physiological signals (Koelstra et al., 2012). For EEG-eye multimodal learning, EmotionMeter showed that EEG and eye movement signals exhibit complementary discriminative properties across emotion categories (Zheng et al., 2019). Early studies mainly explored feature-level and decision-level fusion. For example, Lu et al. (2015) compared multiple fusion strategies for EEG and eye move-

ments, while BDAE learned a shared latent representation through a bimodal deep autoencoder (Liu et al., 2016). Correlation-based methods such as DCCA further aligned latent representations across modalities by maximizing cross-modal correlation (Andrew et al., 2013; Liu et al., 2019). These studies motivated the transition from direct feature combination to learned cross-modal representations. More broadly, deep multimodal learning has progressed from learning shared representations across modalities (Ngiam et al., 2011) to explicitly modeling higher-order interactions (Zadeh et al., 2017; Liu et al., 2018) and directional cross-modal attention (Tsai et al., 2019). MISA further separates modality-invariant and modality-specific factors to preserve both shared and complementary information (Hazarika et al., 2020). In EEG-eye emotion recognition, ETF employs Transformer encoders and attention-based fusion to model complementary information between the two modalities (Wang et al., 2021).

Recent work has increasingly focused on cross-subject generalization and modality heterogeneity. Liu et al. (2022) compared multimodal deep learning models in terms of recognition performance and robustness, showing that different fusion strategies can vary substantially in stability and generalization. DMMR addressed cross-subject EEG emotion recognition through denoising mixed mutual reconstruction for domain-generalizable representation learning (Wang et al., 2024). More closely related to EEG-eye fusion, CMSLNet introduced adaptive cross-modal matching (ACM), attentive low-rank multimodal fusion (Att-LMF), and domain generalization to jointly model modality consistency and cross-modal interactions (Chen et al., 2025). Beyond multimodal fusion, recent EEG methods have also explored structured and discrete representations. MIND-EEG combines multi-granularity brain-region modeling with discrete codebook learning (Zhang et al., 2026), while EEG-SCMM improves cross-corpus generalization through soft contrastive mask modeling (Liu et al., 2025).

These methods primarily focus on task utility and cross-subject generalization, with fused representations typically treated as task-oriented features. Whether cross-modal interactions also preserve or amplify subject-specific information while improving task discrimination remains largely unexplored.

2.2 Identity Leakage and Privacy-Preserving Representation Learning

Physiological and behavioral signals contain stable subject-specific patterns that can support identity inference. In EEG-based brain–computer interfaces, Meng et al. (2023) showed that learned EEG representations may expose user identity and highlighted the need for identity protection. Eye movements exhibit similar properties. Reading scanpaths have been used for biometric identification (Holland & Komogortsev, 2011), and complex eye movement patterns remain discriminative across different environments and stimuli (Holland & Komogortsev, 2013). More recent studies have learned identity representations directly from eye movement sequences using metric learning and deep temporal models (Lohr et al., 2021; Lohr & Komogortsev, 2022). These findings indicate that both modalities considered in our setting may carry identity-relevant information even when identity is unrelated to the downstream emotion task.

Privacy-preserving representation learning aims to retain information useful for a target task while suppressing sensitive attributes. Adversarial learning is a common mechanism for this purpose. The gradient reversal layer (GRL), originally introduced for learning domain-invariant representations (Ganin & Lempitsky, 2015), allows a feature encoder to be optimized against an auxiliary discriminator and has subsequently motivated sensitive-attribute suppression approaches. For EEG privacy protection, ID-RemovalNet explicitly separates task-related and identity-related components and suppresses identity information while preserving downstream decoding utility (Wang et al., 2025). These methods demonstrate that adversarial learning and representation disentanglement can effectively reduce identity information at the unimodal level.

Discrete representation learning provides a complementary mechanism for limiting fine-grained variation in continuous feature spaces. VQ-VAE maps continuous latent representations to entries in a learned discrete codebook (van den Oord et al., 2017), building on classical vector quantization that approximates a continuous space with a finite set of prototypes (Linde et al., 1980). In EEG emotion recognition, MIND-EEG further uses a discrete codebook to learn reusable prototypes of brain-network patterns (Zhang et al., 2026). Although these methods are not designed specifically

for identity privacy, discrete prototype representations provide a useful mechanism for restricting subject-specific variations that are unnecessary for the target task.

However, these studies do not investigate whether residual identity cues in separately protected modalities can be recombined by a learned fusion operator. Moreover, subject invariance optimized for cross-subject generalization does not necessarily imply resistance to post-hoc identity inference. We address this gap by characterizing fusion-induced identity re-leakage and mitigating it at the interaction, continuous-representation, and discrete-prototype levels through privacy-aware low-rank fusion, fusion-level task–identity disentanglement, and a fused privacy codebook.

3 METHOD

We study identity leakage and privacy protection in multimodal emotion recognition from EEG and eye-movement signals. Given paired EEG and eye-movement inputs $\mathbf{X}_{\text{eeg}}$ and $\mathbf{x}_{\text{eye}}$, together with an emotion label y and a subject identity label s, our goal is to learn a fused representation $\mathbf{z}_{\text{f}}$ that retains emotion-discriminative information while suppressing subject-specific information that enables identity inference. Our methodology comprises two components. First, we establish a fusion-level identity re-leakage verification framework to determine whether privacy protection applied independently to each modality remains effective after multimodal fusion. Second, we propose PrivFusionNet, a privacy-preserving multimodal representation learning framework that directly mitigates identity leakage in the fused representation.

3.1 PROBLEM SETUP

A standard multimodal emotion recognition model first encodes the EEG and eye-movement inputs into modality-specific representations:

$$\mathbf{z}_{\text{eeg}} = \mathcal{E}_{\text{eeg}}(\mathbf{X}_{\text{eeg}}), \qquad \mathbf{z}_{\text{eye}} = \mathcal{E}_{\text{eye}}(\mathbf{x}_{\text{eye}}), \tag{1}$$

where $\mathcal{E}_{\text{eeg}}$ and $\mathcal{E}_{\text{eye}}$ denote the EEG and eye-movement encoders, respectively. The resulting representations are combined by a fusion operator $\mathcal{F}$:

$$\mathbf{z}_{\text{f}} = \mathcal{F}\left(\mathbf{z}_{\text{eeg}}, \mathbf{z}_{\text{eye}}\right), \tag{2}$$

where $\mathbf{z}_{\text{f}} \in \mathbb{R}^d$ is the fused representation. Let $C_y : \mathbb{R}^d \to \mathbb{R}^C$ be an emotion classifier that outputs scores over C emotion classes. The predicted emotion label is

$$\hat{y} = \underset{c\in\{1,\ldots,C\}}{\arg\max} \left[C_y(\mathbf{z}_{\text{f}})\right]_c. \tag{3}$$

Although $\mathbf{z}_{\text{f}}$ is optimized for emotion recognition, it may retain subject-specific cues that enable identity inference. We quantify this leakage using user identification accuracy (UIA). Let $C_s : \mathbb{R}^d \to \mathbb{R}^S$ be an identity classifier that outputs scores over S subjects. Given N representations $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^N$ with identity labels $\{s_i\}_{i=1}^N$, the predicted identity and UIA are defined as

$$\hat{s}_i = \underset{j\in\{1,\ldots,S\}}{\arg\max} \left[C_s(\mathbf{z}_i)\right]_j, \qquad \text{UIA}(\mathcal{Z}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\hat{s}_i = s_i\right), \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. A higher UIA indicates that the representation retains more identity-relevant information, whereas uniform random guessing yields an expected UIA of $1/S$. Therefore, a desirable representation should maintain strong emotion-recognition performance while keeping UIA close to the chance level.

3.2 MODALITY-SPECIFIC ENCODERS

The identity re-leakage verification framework and PrivFusionNet employ the same modality-specific encoder architectures, ensuring that their comparison is based on a shared feature-extraction backbone.

For EEG, each input sample is represented as $\mathbf{X}_{\text{eeg}} \in \mathbb{R}^{C\times B}$, where C and B denote the numbers of EEG channels and frequency bands, respectively. The EEG encoder first applies a learnable band

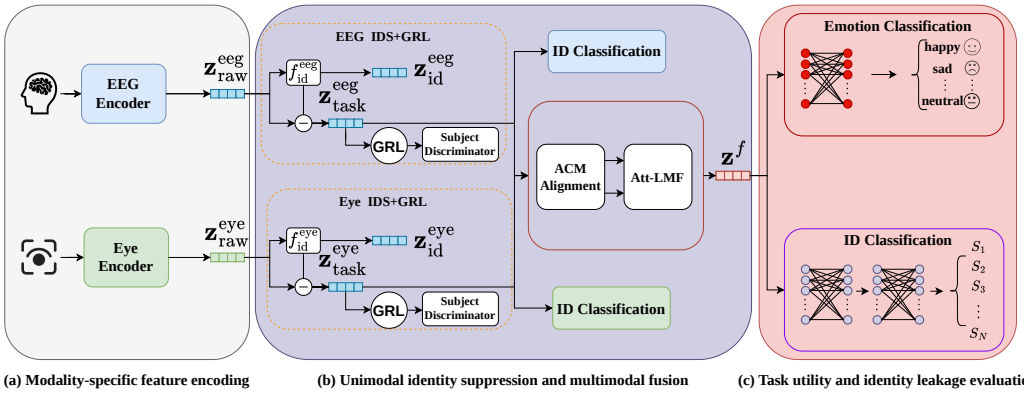


Figure 1: Fusion-level identity re-leakage verification framework. In the original leakage setting, the IDS+GRL branches are disabled and raw unimodal representations are directly fed into the shared fusion module. In the unimodal privacy setting, the IDS+GRL branches are enabled to suppress identity information before fusion. The framework evaluates whether identity leakage reappears in the fused representation after unimodal privacy protection.

gate to adaptively reweight the frequency-band features. It then projects the band features of each channel into a d-dimensional channel token and adds a learnable channel embedding. The resulting tokens are processed by a learned channel-graph mixing layer and a Transformer encoder to capture dependencies among EEG channels. The encoded token sequence is subsequently mean-pooled. To retain complementary global information, the pooled token representation is combined with a flat residual MLP branch and a multiband and multiregion aggregation branch. These features are jointly projected to produce the final EEG representation $\mathbf{z}_{\text{eeg}} \in \mathbb{R}^d$.

For eye movement, each input $\mathbf{x}_{\text{eye}} \in \mathbb{R}^D$ is a pre-extracted statistical feature vector rather than a raw temporal sequence. We therefore use a lightweight residual MLP encoder instead of a temporal sequence model. The encoder first normalizes and nonlinearly projects the input into a d-dimensional latent space and then refines it through stacked residual MLP blocks. The resulting eye-movement representation is denoted by $\mathbf{z}_{\text{eye}} \in \mathbb{R}^d$. Consequently, both modalities are mapped into a shared representation space before multimodal fusion.

3.3 Fusion-Level Identity Re-Leakage Verification

We investigate whether identity suppression performed independently for each modality remains effective after multimodal fusion. As illustrated in Fig. 1, the verification framework compares an original leakage setting with a unimodal privacy setting while keeping the modality-specific encoder architectures and fusion backbone fixed.

Identity leakage in each task-related representation is assessed using an identity classifier. Meanwhile, a subject discriminator connected through a gradient reversal layer encourages the corresponding encoder to suppress subject-discriminative information. The resulting task-related representations are then passed to the same fusion backbone:

$$\mathbf{z}^f = \mathcal{F}_{\text{base}}\left(\mathbf{z}^{\text{eeg}}_{\text{task}}, \mathbf{z}^{\text{eye}}_{\text{task}}\right). \tag{5}$$

Following CMSLNet (Chen et al., 2025), $\mathcal{F}_{\text{base}}$ consists of adaptive cross-modal matching (ACM) followed by attentive low-rank multimodal fusion (Att-LMF). We use UIA to quantify identity leakage in $\mathbf{z}^{\text{eeg}}_{\text{task}}$, $\mathbf{z}^{\text{eye}}_{\text{task}}$, and $\mathbf{z}^f$. We consider fusion-level identity re-leakage to occur when $\mathbf{z}^f$ remains highly identity-identifiable after the protected unimodal representations are fused.

3.4 PrivFusionNet

To directly protect the fused representation, we propose PrivFusionNet, as illustrated in Fig. 2. PrivFusionNet introduces privacy control at both the cross-modal interaction stage and the post-

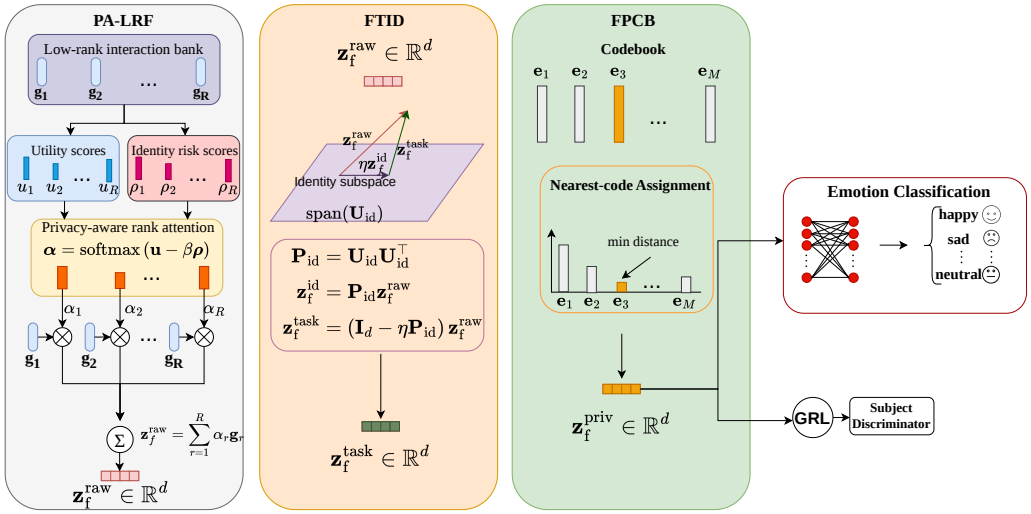


Figure 2: Core privacy-processing pipeline of PrivFusionNet. The front-end components already shown in Fig. 1 are omitted. From left to right, PA-LRF, FTID, and FPCB suppress identity-sensitive interactions, attenuate the fused identity component, and quantize the task-related representation. The resulting privacy code supports emotion classification and GRL-based subject adversarial training.

fusion representation stage. It consists of three key modules: privacy-aware low-rank fusion (PA-LRF), fusion-level task–identity disentanglement (FTID), and the fused privacy codebook (FPCB).

3.4.1 Privacy-Aware Low-Rank Fusion

Low-rank multimodal fusion captures higher-order cross-modal interactions with a compact parameterization (Liu et al., 2018). Given the modality-specific representations $\mathbf{z}^{\text{eeg}}$ and $\mathbf{z}^{\text{eye}}$, PA-LRF constructs R rank-specific interaction components:

$$\mathbf{g}_r = \phi_r(\mathbf{z}^{\text{eeg}}, \mathbf{z}^{\text{eye}}) \in \mathbb{R}^d, \qquad r = 1, \ldots, R, \tag{6}$$

where $\phi_r(\cdot,\cdot)$ denotes the r-th low-rank cross-modal interaction function.

Standard attentive low-rank fusion primarily weights interaction components according to their task utility. However, an interaction component that benefits emotion recognition may simultaneously preserve subject-specific information. We therefore associate each component $\mathbf{g}_r$ with a task-utility score u_r and an identity-risk score ρ_r. The corresponding privacy-aware attention weight is defined as

$$\alpha_r = \frac{\exp(u_r - \beta\rho_r)}{\sum_{j=1}^{R} \exp(u_j - \beta\rho_j)}, \tag{7}$$

where $\beta \geq 0$ controls the strength of identity-risk suppression. Equivalently, the complete attention vector can be written as

$$\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{u} - \beta\boldsymbol{\rho})\,. \tag{8}$$

The rank-specific components are then aggregated to obtain the raw fused representation:

$$\mathbf{z}_f^{\text{raw}} = \sum_{r=1}^{R} \alpha_r \mathbf{g}_r \in \mathbb{R}^d. \tag{9}$$

By jointly considering task utility and identity risk, PA-LRF assigns larger weights to cross-modal interactions that are informative for emotion recognition while suppressing interactions that carry stronger subject-specific cues.

3.4.2 FUSION-LEVEL TASK–IDENTITY DISENTANGLEMENT

Although PA-LRF suppresses identity-risky interactions during fusion, the resulting continuous fused representation may still retain residual subject-specific information. FTID further disentangles task-relevant and identity-related information by learning a low-dimensional identity subspace in the fused representation space.

Let $\mathbf{z}_f^{\mathrm{raw}} \in \mathbb{R}^d$ denote the raw fused representation. FTID learns an identity basis

$$\mathbf{U}_{\mathrm{id}} \in \mathbb{R}^{d\times k}, \qquad k < d, \tag{10}$$

whose columns span the identity-related subspace. With approximately orthonormal basis vectors, the corresponding projection matrix is

$$\mathbf{P}_{\mathrm{id}} = \mathbf{U}_{\mathrm{id}}\mathbf{U}_{\mathrm{id}}^{\top} \in \mathbb{R}^{d\times d}. \tag{11}$$

The identity-related component of the fused representation is obtained as

$$\mathbf{z}_f^{\mathrm{id}} = \mathbf{P}_{\mathrm{id}}\mathbf{z}_f^{\mathrm{raw}}. \tag{12}$$

FTID then applies controllable projection suppression to obtain the task-related representation:

$$\mathbf{z}_f^{\mathrm{task}} = (\mathbf{I}_d - \eta\mathbf{P}_{\mathrm{id}})\,\mathbf{z}_f^{\mathrm{raw}}, \qquad \eta \in [0,1], \tag{13}$$

where η controls the suppression strength. When $\eta = 0$, the fused representation remains unchanged; when $\eta = 1$, the projected identity component is fully removed. Intermediate values provide partial suppression, reducing identity information while avoiding excessive removal of task-relevant cues.

3.4.3 FUSED PRIVACY CODEBOOK

Even after FTID, continuous fused representations may retain fine-grained subject-specific variations. FPCB reduces such variations by quantizing the task-related fused representation into a finite set of subject-shared, task-oriented prototypes. We define a learnable codebook $\mathcal{E}_{\mathrm{cb}} = \{\mathbf{e}_m\}_{m=1}^{M}$, where M is the codebook size and $\mathbf{e}_m \in \mathbb{R}^d$.

Code assignment is performed using ℓ_2-normalized representations to reduce sensitivity to feature magnitude. Let $\overline{\mathbf{z}}_f^{\mathrm{task}}$ and $\overline{\mathbf{e}}_m$ denote the normalized task representation and code vectors, respectively. The nearest code is selected as

$$m^* = \underset{m\in\{1,\dots,M\}}{\arg\min} \left\|\overline{\mathbf{z}}_f^{\mathrm{task}} - \overline{\mathbf{e}}_m\right\|_2^2. \tag{14}$$

We employ a straight-through estimator to obtain the protected fused representation:

$$\mathbf{z}_f^{\mathrm{priv}} = \mathbf{z}_f^{\mathrm{task}} + \mathrm{sg}\left(\mathbf{e}_{m^*} - \mathbf{z}_f^{\mathrm{task}}\right), \tag{15}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Thus, the forward representation equals the selected code vector, while gradients from downstream objectives can still propagate to the preceding modules.

During training, we additionally apply a class-conditional decorrelation regularizer between $\mathbf{z}_f^{\mathrm{priv}}$ and $\mathbf{z}_f^{\mathrm{id}}$ to reduce residual identity dependence without penalizing task-discriminative variation between classes. The complete training objective is provided in Appendix A.

The quantizer is optimized using

$$\mathcal{L}_{\mathrm{cb}} = \left\|\mathbf{z}_f^{\mathrm{task}} - \mathrm{sg}(\mathbf{e}_{m^*})\right\|_2^2 + \beta_{\mathrm{cb}}\left\|\mathbf{e}_{m^*} - \mathrm{sg}(\mathbf{z}_f^{\mathrm{task}})\right\|_2^2, \tag{16}$$

where β_{cb} balances the two quantization terms.

We additionally regularize code utilization using soft assignments derived from the same normalized distances. Let $\mathbf{p}$ denote the average code-usage distribution in a mini-batch and $\mathbf{p}^{(s)}$ the corresponding distribution for subject s. We define

$$\mathcal{L}_{\mathrm{usage}} = 1 - \frac{H(\mathbf{p})}{\log M}, \qquad \mathcal{L}_{\mathrm{cb\text{-}sub}} = \frac{1}{M|\mathcal{S}_{\mathcal{B}}|}\sum_{s\in\mathcal{S}_{\mathcal{B}}}\left\|\mathbf{p}^{(s)} - \mathrm{sg}(\mathbf{p})\right\|_2^2, \tag{17}$$

Table 1: Task utility and identity leakage at different representation levels.

		EEG (%)	**Eye** (%)	**Fused** (%)
Method	**BCA** (%) ↑	**UIA** ↓	**UIA** ↓	**UIA** ↓
SEED-IV				
Original Leakage	73.14	43.98	41.47	25.33
Unimodal IDS Baseline	71.51	4.90	7.61	28.30
PrivFusionNet	71.49	55.00	43.48	**7.57**
SEED-V				
Original Leakage	65.39	74.13	60.44	46.38
Unimodal IDS Baseline	65.04	7.90	7.64	47.35
PrivFusionNet	66.35	81.95	59.82	**6.11**

where

$$H(\mathbf{p}) = -\sum_{m=1}^{M} p_m \log(p_m + \epsilon), \tag{18}$$

and $\mathcal{S}_{\mathcal{B}}$ is the set of subjects represented in the current mini-batch. The first term prevents codebook collapse by encouraging diverse code utilization, whereas the second prevents code-selection preferences from becoming subject-specific.

The emotion classifier operates on the quantized representation and predicts the emotion label as

$$\widehat{y} = \underset{c\in\{1,\ldots,C\}}{\arg\max} \left[C_y \left(\mathbf{z}_f^{\text{priv}} \right) \right]_c .$$

During training, $\mathbf{z}_f^{\text{priv}}$ is also passed to a subject discriminator through a gradient reversal layer, further reducing subject-identifiable information. The discriminator is removed at inference time. Consequently, FPCB learns subject-shared discrete prototypes that preserve emotion-related information while suppressing fine-grained identity cues.

4 Results

We organize the experiments around two questions. First, can modality-wise identity suppression prevent identity re-leakage after multimodal fusion? Second, can PrivFusionNet reduce identity leakage in fused representations while preserving cross-subject emotion recognition performance?

4.1 Fusion-level identity re-leakage

We first investigate whether unimodal identity suppression is sufficient to protect the representation produced by subsequent multimodal fusion. We evaluate emotion recognition using balanced classification accuracy (BCA) and identity leakage using user identification accuracy (UIA). Higher BCA indicates better task utility, whereas lower UIA indicates less recoverable identity information.

Unimodal identity suppression does not prevent fusion-level re-leakage. On SEED-IV, the unimodal IDS baseline reduces EEG UIA from 43.98% to 4.90% and eye-movement UIA from 41.47% to 7.61%. Nevertheless, UIA for the fused representation increases from 25.33% to 28.30%. The same behavior is observed on SEED-V: EEG and eye-movement UIA decrease from 74.13% and 60.44% to 7.90% and 7.64%, respectively, whereas UIA for the fused representation increases from 46.38% to 47.35%.

These results show that privacy at the unimodal level is not preserved automatically under multimodal fusion. Even when each modality is weakly identifiable in isolation, the fusion operator can combine complementary residual cues and reconstruct a representation that is informative of subject identity. Fusion-level privacy is therefore not a simple average of the privacy properties of the input modalities and must be addressed explicitly at and after the fusion operation.

Table 2: Cross-subject emotion recognition accuracy under LOSO. Results are aggregated over all 45 subject–session folds on SEED-IV and all 48 folds on SEED-V.

Method	SEED-IV ACC (%) ↑	SEED-V ACC (%) ↑
Concatenation	43.82	42.85
DLF-SUM	63.67	58.00
DLF-MAX	61.43	55.15
FLF	57.77	52.28
DCCA-fusion	78.18	82.72
ETF	79.53	81.02
CMSLNet	83.15	87.32
PrivFusionNet	**84.23**	**88.30**

PrivFusionNet suppresses identity information in the fused representation. PrivFusionNet reduces UIA for the fused representation to 7.57% on SEED-IV and 6.11% on SEED-V. Relative to the unprotected fusion model, these correspond to absolute reductions of 17.76 and 40.27 percentage points, respectively. The resulting values are close to the random-identification levels of 6.67% for SEED-IV and 6.25% for SEED-V.

This privacy improvement incurs only a small utility change. On SEED-IV, PrivFusionNet obtains a BCA of 71.49%, compared with 73.14% for unprotected fusion. On SEED-V, it improves BCA from 65.39% to 66.35%. The results suggest that suppressing identity cues in the fused space does not necessarily require discarding the cross-modal information needed for emotion recognition.

4.2 Cross-subject emotion recognition

We next evaluate whether fusion-level privacy protection compromises generalization to unseen subjects. We use the strict source-only leave-one-subject-out (LOSO) protocol. LOSO is conducted independently within each of the three sessions, after which all subject–session folds are pooled to compute the overall result. This produces 45 folds for SEED-IV and 48 folds for SEED-V. The held-out subject is not used for training, validation, or hyperparameter selection.

As shown in Table 2, simple feature concatenation performs substantially worse than methods that explicitly model cross-modal interactions. Decision-level fusion methods improve over concatenation on SEED-IV, but their ability to capture fine-grained dependencies between EEG and eye movement is limited. Representation-level fusion methods, including DCCA-fusion, ETF, and CMSLNet, provide stronger cross-subject performance.

PrivFusionNet achieves an ACC of 84.23% on SEED-IV and 88.30% on SEED-V. It exceeds the strongest listed baseline, CMSLNet, by 1.08 and 0.98 percentage points, respectively. These results indicate that the proposed privacy constraints do not prevent the model from learning emotion-discriminative cross-modal representations. Instead, explicitly controlling identity information during fusion can coexist with strong generalization to unseen subjects.

5 Conclusion

We identify fusion-induced identity re-leakage in EEG–eye emotion recognition, showing that unimodal privacy protection does not necessarily safeguard fused representations. We propose PrivFusionNet, which combines privacy-aware low-rank fusion, task–identity disentanglement, and discrete prototype learning to mitigate this risk. Experiments on SEED-IV and SEED-V demonstrate substantially reduced identity leakage while maintaining competitive cross-subject emotion recognition performance. These findings highlight the need to protect and evaluate identity privacy directly at the fusion level.

AI use statement

In this work, we used large language models (LLMs) to assist with translation and language editing to improve the expression and readability of the manuscript. We also used LLMs to assist with

the implementation and reproduction of four conventional baseline methods: Concatenation, DLF-SUM, DLF-MAX, and FLF. All AI-assisted textual content and code were reviewed, revised, and verified by two authors, and the generated code was checked for correctness before being used in our experiments. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses publicly available EEG and eye movement datasets and does not involve the collection of new human-subject data. Our study focuses on identity leakage from multimodal physiological representations and is intended to improve privacy protection in multimodal learning systems. We do not attempt to identify real-world individuals beyond the subject labels provided by the datasets. The identity inference experiments are used only to quantify privacy leakage under controlled evaluation settings. We acknowledge that physiological and behavioral signals may contain sensitive personal information, and practical deployment of such systems should follow appropriate consent, data protection, and access-control practices.

REFERENCES

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1247–1255. PMLR, 2013.

Prithila Angkan, Behnam Behinaein, Zunayed Mahmud, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. Multimodal brain–computer interface for in-vehicle driver cognitive load measurement: Dataset and baselines. *IEEE Transactions on Intelligent Transportation Systems*, 25(6):5949–5964, 2024. doi: 10.1109/TITS.2023.3345846.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Chuangquan Chen, Zhencheng Li, Kit Ian Kou, Jie Du, Chen Li, Hongtao Wang, and Chi-Man Vong. Comprehensive multisource learning network for cross-subject multimodal emotion recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9(1):365–380, 2025. doi: 10.1109/TETCI.2024.3406422.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1–10, 2018. doi: 10.18653/v1/D18-1001.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1180–1189, 2015.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020. doi: 10.1145/3394171.3413678.

Corey Holland and Oleg V. Komogortsev. Biometric identification via eye movement scanpaths in reading. In *2011 International Joint Conference on Biometrics (IJCB)*, pp. 1–8. IEEE, 2011. doi: 10.1109/IJCB.2011.6117536.

Corey D. Holland and Oleg V. Komogortsev. Complex eye movement pattern biometrics: The effects of environment and stimulus. *IEEE Transactions on Information Forensics and Security*, 8(12):2115–2126, 2013. doi: 10.1109/TIFS.2013.2285884.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. doi: 10.1109/T-AFFC.2011.15.

Yoseph Linde, Andres Buzo, and Robert M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28(1):84–95, 1980. doi: 10.1109/TCOM.1980.1094577.

Qile Liu, Weishan Ye, Lingli Zhang, and Zhen Liang. EEG-SCMM: Soft contrastive masked modeling for cross-corpus EEG-based emotion recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5834–5842, 2025. doi: 10.1145/3746027.3755798.

Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal deep learning. In *Proceedings of the International Conference on Neural Information Processing*, pp. 521–529. Springer, 2016. doi: 10.1007/978-3-319-46672-9_58.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using deep canonical correlation analysis. *arXiv preprint arXiv:1908.05349*, 2019.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022. doi: 10.1109/TCDS.2021.3071170.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2247–2256, 2018. doi: 10.18653/v1/P18-1209.

Dillon Lohr and Oleg V. Komogortsev. Eye know you too: Toward viable end-to-end eye movement biometrics for user authentication. *IEEE Transactions on Information Forensics and Security*, 17: 3151–3164, 2022. doi: 10.1109/TIFS.2022.3201369.

Dillon Lohr, Henry Griffith, and Oleg V. Komogortsev. Eye know you: Metric learning for end-to-end biometric authentication using eye movements from a longitudinal dataset. *arXiv preprint arXiv:2104.10489*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and EEG to enhance emotion recognition. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pp. 1170–1176, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3384–3393. PMLR, 2018.

Lubin Meng, Xue Jiang, Jian Huang, Wei Li, Hanbin Luo, and Dongrui Wu. User identity protection in EEG-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3576–3586, 2023. doi: 10.1109/TNSRE.2023.3310883.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 689–696, 2011.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Huabin Wang, Jie Ruan, Cunhang Fan, Yingfan Cheng, and Zhao Lv. ID-RemovalNet: Identity removal network for EEG privacy protection with enhancing decoding tasks. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4209–4217, 2025. doi: 10.24963/ijcai.2025/469.

Yiming Wang, Bin Zhang, and Yujiao Tang. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 628–636, 2024. doi: 10.1609/aaai.v38i1.27819.

Yiting Wang, Wei-Bang Jiang, Rui Li, and Bao-Liang Lu. Emotion transformer fusion: Complementary representation properties of EEG and eye movements on recognizing anger and surprise. In *2021 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1575–1578, 2021. doi: 10.1109/BIBM52615.2021.9669556.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017. doi: 10.18653/v1/D17-1115.

Yuzhe Zhang, Chengxi Xie, Huan Liu, Yuhan Shi, Guanjian Liu, and Dalin Zhang. MIND-EEG: Multi-granularity integration network with discrete codebook for EEG-based emotion recognition. *IEEE Transactions on Affective Computing*, 17(1):147–161, 2026. doi: 10.1109/TAFFC.2025.3608571.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49 (3):1110–1122, 2019. doi: 10.1109/TCYB.2018.2797176.

APPENDIX

A TRAINING OBJECTIVE

PrivFusionNet is trained using task-utility objectives, fusion-level privacy objectives, and representation regularization. Let

$$\mathcal{B} = \{(\mathbf{X}_{\mathrm{eeg},i}, \mathbf{x}_{\mathrm{eye},i}, y_i, s_i)\}_{i=1}^{N_b}$$

denote a mini-batch of size N_b, where y_i and s_i are the task and subject labels, respectively. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \mathcal{L}_{\mathrm{priv}} + \mathcal{L}_{\mathrm{repr}}. \tag{19}$$

Task-utility objective. The task-utility objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{task}} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{aux}}^{\mathrm{eeg}}\mathcal{L}_{\mathrm{cls}}^{\mathrm{eeg}} + \lambda_{\mathrm{aux}}^{\mathrm{eye}}\mathcal{L}_{\mathrm{cls}}^{\mathrm{eye}} \\ + \lambda_{\mathrm{acm}}\mathcal{L}_{\mathrm{acm}} + \lambda_{\mathrm{tc}}\mathcal{L}_{\mathrm{tc}} + \lambda_{\mathrm{dfs}}\mathcal{L}_{\mathrm{dfs}}.\end{aligned} \tag{20}$$

The primary task-classification loss is computed from the protected fused representation:

$$\mathcal{L}_{\mathrm{cls}} = \frac{1}{N_b}\sum_{i=1}^{N_b} \mathrm{CE}\left(C_y(\mathbf{z}_{f,i}^{\mathrm{priv}}), y_i\right), \tag{21}$$

where C_y denotes the task classifier and CE denotes cross-entropy.

To provide direct task supervision for both modality-specific encoders, we additionally use

$$\mathcal{L}_{\mathrm{cls}}^{\nu} = \frac{1}{N_b}\sum_{i=1}^{N_b} \mathrm{CE}\left(C_y^{\nu}(\mathbf{z}_i^{\nu}), y_i\right), \qquad \nu \in \{\mathrm{eeg}, \mathrm{eye}\}. \tag{22}$$

These auxiliary losses encourage each modality to retain task-discriminative information and prevent the fused model from depending excessively on a single modality.

The adaptive cross-modal matching loss measures the distance between the modality-specific representations:

$$d_i = \|\mathbf{z}_i^{\mathrm{eeg}} - \mathbf{z}_i^{\mathrm{eye}}\|_2, \qquad \mathcal{L}_{\mathrm{acm}} = \frac{1}{N_b}\sum_{i=1}^{N_b} h_\delta(d_i), \tag{23}$$

where

$$h_\delta(d) = \begin{cases} \frac{1}{2}d^2, & d \le \delta, \\ \delta d - \frac{1}{2}\delta^2, & d > \delta. \end{cases} \tag{24}$$

This term encourages task-relevant consistency between EEG and eye-movement representations while reducing the influence of excessively large cross-modal discrepancies.

We further apply supervised task contrastive learning to the protected fused representation. For anchor i, its positive set is

$$\mathcal{P}(i) = \{j \neq i \mid y_j = y_i,\ s_j \neq s_i\}. \tag{25}$$

Using the normalized representation $\overline{\mathbf{z}}_{f,i}^{\mathrm{priv}}$, the contrastive loss is

$$\mathcal{L}_{\mathrm{tc}} = -\frac{1}{|\mathcal{I}_{\mathcal{B}}|}\sum_{i\in\mathcal{I}_{\mathcal{B}}}\frac{1}{|\mathcal{P}(i)|}\sum_{j\in\mathcal{P}(i)}\log\frac{\exp\left(\overline{\mathbf{z}}_{f,i}^{\mathrm{priv}\,\top}\overline{\mathbf{z}}_{f,j}^{\mathrm{priv}}/\tau\right)}{\sum_{a\neq i}\exp\left(\overline{\mathbf{z}}_{f,i}^{\mathrm{priv}\,\top}\overline{\mathbf{z}}_{f,a}^{\mathrm{priv}}/\tau\right)}, \tag{26}$$

where $\mathcal{I}_{\mathcal{B}} = \{i \mid |\mathcal{P}(i)| > 0\}$ and τ is the temperature. Restricting positive pairs to samples from different subjects encourages task-level clustering without reinforcing subject-specific similarity.

The dominant-feature suppression loss prevents the task classifier from relying excessively on a small number of high-response feature dimensions. Let $\mathcal{T}_i$ contain the largest-magnitude dimensions of $\mathbf{z}_{f,i}^{\mathrm{priv}}$. We construct a binary mask $\mathbf{b}_i$ that sets these dimensions to zero and define

$$\widetilde{\mathbf{z}}_{f,i}^{\mathrm{priv}} = \mathbf{b}_i \odot \mathbf{z}_{f,i}^{\mathrm{priv}}, \qquad \mathcal{L}_{\mathrm{dfs}} = \frac{1}{N_b}\sum_{i=1}^{N_b}\mathrm{CE}\left(C_y(\widetilde{\mathbf{z}}_{f,i}^{\mathrm{priv}}), y_i\right). \tag{27}$$

Fusion-level privacy objective. The fusion-level privacy objective is

$$\mathcal{L}_{\text{priv}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{cb-sub}}\mathcal{L}_{\text{cb-sub}}. \tag{28}$$

The fused subject-adversarial loss is

$$\mathcal{L}_{\text{adv}} = \frac{1}{N_b}\sum_{i=1}^{N_b} \text{CE}\left(D_s\left(\text{GRL}_\gamma(\mathbf{z}_{f,i}^{\text{priv}})\right), s_i\right), \tag{29}$$

where D_s is the subject discriminator. The gradient reversal layer is the identity operation during the forward pass but multiplies the gradient passed to preceding modules by $-\gamma$. Consequently, the discriminator learns to identify subjects, whereas the representation network learns to suppress subject-identifiable information.

To ensure that the identity subspace learned by FTID captures subject-related information, we define

$$\mathcal{L}_{\text{id}} = \frac{1}{N_b}\sum_{i=1}^{N_b} \text{CE}\left(D_{\text{id}}(\mathbf{z}_{f,i}^{\text{id}}), s_i\right). \tag{30}$$

This loss concentrates subject-discriminative information in $\mathbf{z}_f^{\text{id}}$, allowing FTID to suppress it along the task pathway.

PA-LRF utility and identity-risk estimation. For sample i, let $\mathbf{g}_{i,r} \in \mathbb{R}^d$ denote the r-th low-rank interaction component, where $r \in \{1, \ldots, R\}$. PA-LRF first normalizes each interaction component as

$$\overline{\mathbf{g}}_{i,r} = \frac{\mathbf{g}_{i,r}}{\|\mathbf{g}_{i,r}\|_2 + \epsilon}, \tag{31}$$

where $\epsilon > 0$ is a small constant for numerical stability.

Let $\mathbf{q} \in \mathbb{R}^d$ be a learnable utility query and

$$\overline{\mathbf{q}} = \frac{\mathbf{q}}{\|\mathbf{q}\|_2 + \epsilon} \tag{32}$$

its normalized form. The initial utility score of the r-th interaction component is

$$\widetilde{u}_{i,r} = \overline{\mathbf{g}}_{i,r}^\top \overline{\mathbf{q}}. \tag{33}$$

Writing $\widetilde{\mathbf{u}}_i = [\widetilde{u}_{i,1}, \ldots, \widetilde{u}_{i,R}]^\top$, the final task-utility score is

$$u_{i,r} = \omega_r \left[\text{LN}\left(\widetilde{\mathbf{u}}_i\right)\right]_r + \delta_r, \tag{34}$$

where $\text{LN}(\cdot)$ applies layer normalization across the R interaction scores, and ω_r and δ_r are learnable rank-specific scale and bias parameters.

Identity risk is estimated directly at the interaction-component level. First, a learnable local-risk scorer produces

$$\ell_{i,r} = \sigma\left(h_\psi\left(\overline{\mathbf{g}}_{i,r}\right)\right), \tag{35}$$

where $h_\psi : \mathbb{R}^d \to \mathbb{R}$ is a two-layer MLP and $\sigma(\cdot)$ denotes the sigmoid function.

In parallel, a rank-wise subject classifier $D_{\text{rank}} : \mathbb{R}^d \to \mathbb{R}^S$ estimates how confidently the interaction component reveals the identity of one of the S training subjects:

$$\boldsymbol{\pi}_{i,r}^{\text{rank}} = \text{softmax}\left(D_{\text{rank}}\left(\text{sg}\left(\overline{\mathbf{g}}_{i,r}\right)\right)\right), \qquad \kappa_{i,r}^{\text{rank}} = \max_{s\in\{1,\ldots,S\}} \left[\boldsymbol{\pi}_{i,r}^{\text{rank}}\right]_s, \tag{36}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. A larger $\kappa_{i,r}^{\text{rank}}$ indicates that the corresponding interaction component permits more confident subject identification.

The final identity-risk score combines the learned local estimate and the supervised subject-classification confidence:

$$\rho_{i,r} = \frac{1}{2}\ell_{i,r} + \frac{1}{2}\,\text{sg}\left(\kappa_{i,r}^{\text{rank}}\right). \tag{37}$$

The resulting utility and identity-risk scores are substituted into the privacy-aware attention weights defined in Eq. (7). Thus, interaction components with high task utility and low identity risk receive larger fusion weights.

The rank-wise subject classifier is trained using

$$\mathcal{L}_{\text{rank}} = \frac{1}{N_b R} \sum_{i=1}^{N_b} \sum_{r=1}^{R} \text{CE}\left(D_{\text{rank}}\left(\text{sg}\left(\overline{\mathbf{g}}_{i,r}\right)\right), s_i\right), \tag{38}$$

where N_b is the mini-batch size and s_i is the subject label of sample i. The stop-gradient operator prevents $\mathcal{L}_{\text{rank}}$ from directly modifying the interaction components. It also prevents the privacy-aware attention objective from changing D_{rank} through $\kappa_{i,r}^{\text{rank}}$. The utility scorer and the local-risk scorer are optimized jointly with the downstream task and privacy objectives through the attention weights.

For FPCB, let

$$q_{im} = \frac{\exp\left(-\left\|\overline{\mathbf{z}}_{f,i}^{\text{task}} - \overline{\mathbf{e}}_m\right\|_2^2\right)}{\sum_{j=1}^{M} \exp\left(-\left\|\overline{\mathbf{z}}_{f,i}^{\text{task}} - \overline{\mathbf{e}}_j\right\|_2^2\right)} \tag{39}$$

denote the soft assignment probability of sample i to code m. The global and subject-conditional code-usage distributions are

$$p_m = \frac{1}{N_b} \sum_{i=1}^{N_b} q_{im}, \qquad p_m^{(s)} = \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} q_{im}, \tag{40}$$

where $\mathcal{B}_s$ contains samples from subject s. We define

$$\mathcal{L}_{\text{cb-sub}} = \frac{1}{M|\mathcal{S}_{\mathcal{B}}|} \sum_{s \in \mathcal{S}_{\mathcal{B}}} \left\|\mathbf{p}^{(s)} - \text{sg}(\mathbf{p})\right\|_2^2, \tag{41}$$

where $\mathcal{S}_{\mathcal{B}}$ is the set of subjects represented in the current mini-batch. This term discourages different subjects from developing distinct code-selection preferences.

Representation regularization. The representation-regularization objective is

$$\mathcal{L}_{\text{repr}} = \lambda_{\text{basis}}\mathcal{L}_{\text{basis}} + \lambda_{\text{ccorr}}\mathcal{L}_{\text{ccorr}} + \lambda_{\text{cb}}\mathcal{L}_{\text{cb}} + \lambda_{\text{usage}}\mathcal{L}_{\text{usage}}. \tag{42}$$

The FTID identity basis is regularized toward orthonormality using

$$\mathcal{L}_{\text{basis}} = \left\|\mathbf{U}_{\text{id}}^{\top}\mathbf{U}_{\text{id}} - \mathbf{I}_k\right\|_F^2. \tag{43}$$

Class-conditional decorrelation reduces dependence between the identity component and the final protected representation:

$$\mathcal{L}_{\text{ccorr}} = \frac{1}{|\mathcal{C}_{\mathcal{B}}|d^2} \sum_{c \in \mathcal{C}_{\mathcal{B}}} \left\|\text{Corr}\left(\mathbf{z}_f^{\text{priv}}, \mathbf{z}_f^{\text{id}} \mid y = c\right)\right\|_F^2, \tag{44}$$

where $\mathcal{C}_{\mathcal{B}}$ denotes the set of task classes represented by at least two samples in the current mini-batch.

Let m_i^* denote the nearest code selected for sample i. The codebook quantization loss is

$$\begin{aligned}\mathcal{L}_{\text{cb}} = \frac{1}{N_b} \sum_{i=1}^{N_b} \Big[&\left\|\mathbf{z}_{f,i}^{\text{task}} - \text{sg}(\mathbf{e}_{m_i^*})\right\|_2^2 \\ &+ \beta_{\text{cb}} \left\|\mathbf{e}_{m_i^*} - \text{sg}(\mathbf{z}_{f,i}^{\text{task}})\right\|_2^2 \Big].\end{aligned} \tag{45}$$

Finally, codebook collapse is discouraged by maximizing the normalized entropy of the average code-usage distribution:

$$H(\mathbf{p}) = -\sum_{m=1}^{M} p_m \log(p_m + \epsilon), \qquad \mathcal{L}_{\text{usage}} = 1 - \frac{H(\mathbf{p})}{\log M}. \tag{46}$$

Table 3: Dataset statistics. EEG inputs contain differential entropy features from 62 channels and five frequency bands.

Dataset	Subjects	Sessions	Trials/session	Classes	EEG	Eye
SEED-IV	15	3	24	4	62×5	31
SEED-V	16	3	15	5	62×5	33

Minimizing $\mathcal{L}_{\text{usage}}$ increases code-usage entropy and encourages diverse utilization of the codebook. All loss weights are nonnegative hyperparameters, and their values are reported in the implementation details. The GRL-connected subject discriminator is used only during training, while task prediction is performed from $\mathbf{z}_f^{\text{priv}}$ at inference time.

B Experimental Setup

We conduct two complementary evaluations. First, we examine whether identity information remains recoverable from fused representations after unimodal privacy protection. Second, we evaluate cross-subject emotion recognition using leave-one-subject-out (LOSO) evaluation. Together, these experiments characterize the trade-off between emotion-recognition utility and identity privacy.

B.1 Datasets and Preprocessing

We evaluate PrivFusionNet on SEED-IV and SEED-V, two public EEG–eye movement emotion-recognition datasets containing temporally aligned 4-second feature windows. Their statistics are summarized in Table 3.

SEED-IV. SEED-IV contains 15 subjects, each recorded in three sessions conducted on different days (Zheng et al., 2019). Each session contains 24 film trials covering neutral, sad, fear, and happy emotions. Each EEG window contains differential entropy features from 62 channels and five frequency bands, and the aligned eye-movement window contains 31 statistical features.

SEED-V. SEED-V contains 16 subjects and three sessions per subject, with 15 film trials in each session (Liu et al., 2022). It includes disgust, fear, sad, neutral, and happy emotions. Each EEG window has the same 62×5 differential entropy representation as SEED-IV, while the eye-movement representation contains 33 features.

Within each fold, missing-value imputation and normalization statistics are estimated exclusively from the training partition and then applied unchanged to the held-out partition. Missing EEG and eye-movement values are imputed using feature-wise medians estimated from the training partition. EEG features are standardized using the training mean and standard deviation, whereas eye-movement features are centered by the training median and scaled by the interquartile range. The same preprocessing is applied to PrivFusionNet and all locally reproduced baselines. No target-domain alignment or target-derived normalization is used.

B.2 Evaluation Protocols

Cross-session identity re-leakage evaluation. We adopt a leave-one-session-out protocol to evaluate identity leakage under session shift. In each fold, two sessions are used for training and the remaining session is held out for evaluation. All subjects appear in both partitions, while samples from the held-out session are excluded from model optimization.

The representation model is trained for 80 epochs, and the best-performing model is selected. Identity leakage is evaluated using an MLP identity classifier trained on representations from the two training sessions and tested on the held-out session.

We evaluate identity leakage from the EEG representation $\mathbf{z}^{\text{eeg}}$, eye-movement representation $\mathbf{z}^{\text{eye}}$, and fused representation $\mathbf{z}_f$. The original leakage model, unimodal privacy baseline, and Priv-

FusionNet use identical session splits and identity-classifier settings. We report mean emotion-recognition BCA and UIA over the three folds.

Cross-subject LOSO evaluation. We perform LOSO evaluation independently within each session. For each subject, the remaining subjects from the same session are used for training. This produces 45 subject–session folds for SEED-IV and 48 folds for SEED-V.

Each model is trained for 80 epochs, and the best-performing model is selected. We report mean window-level ACC over all subject–session folds.

B.3 Evaluation Metrics

Emotion recognition is evaluated using balanced classification accuracy (BCA) in the cross-session experiments and accuracy (ACC) in the LOSO experiments. Let N denote the number of evaluation samples, y_i the ground-truth emotion label of sample i, and $\hat{y}_i$ its predicted label. ACC is defined as

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i]\,, \tag{47}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. Let C denote the number of emotion classes. BCA is the macro-average of the class-wise recall:

$$\mathrm{BCA} = \frac{1}{C}\sum_{c=1}^{C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \tag{48}$$

where TP_c and FN_c denote the numbers of true-positive and false-negative samples for emotion class c, respectively. In contrast to ACC, BCA assigns equal weight to every emotion class and is therefore less affected by class imbalance.

Identity leakage is measured using user identification accuracy (UIA). Given the representations $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^{N}$, their subject labels $\{s_i\}_{i=1}^{N}$, and an identity classifier a_ϕ, UIA is defined as

$$\mathrm{UIA}(\mathcal{Z}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[a_\phi(\mathbf{z}_i) = s_i]\,. \tag{49}$$

Higher UIA indicates stronger identity leakage. The chance levels are 6.67% for SEED-IV and 6.25% for SEED-V. We report the mean results over all evaluation folds.

B.4 Baselines

We compare PrivFusionNet with conventional multimodal fusion baselines and recent cross-subject methods.

We locally reproduce Concatenation, feature-level fusion (FLF), probability-sum decision fusion (DLF-SUM), and maximum-confidence decision fusion (DLF-MAX). These baselines use the same input features, preprocessing, subject–session splits, and evaluation metrics as PrivFusionNet.

Concatenation directly combines EEG and eye-movement features before classification. FLF follows the feature-level fusion setting of Lu et al. (2015). DLF-SUM and DLF-MAX train separate EEG and eye-movement classifiers and combine their output probabilities by summation or element-wise maximum, respectively. The locally reproduced conventional fusion baselines use class-balanced linear support vector machines.

DCCA-fusion and ETF are based on deep canonical correlation analysis and Emotion Transformer Fusion, respectively (Andrew et al., 2013; Wang et al., 2021). For DCCA-fusion, ETF, and CM-SLNet, we use the numerical results reported by Chen et al. (2025), which evaluates these methods on SEED-IV and SEED-V under a compatible within-session LOSO protocol.

Table 4: Main hyperparameters used for PrivFusionNet.

Component	Setting
Optimization	AdamW, 80 epochs, batch size 256, learning rate 3×10^{-4}, weight decay 10^{-4}
PA-LRF	$R = 16$, $\beta = 0.5$, $\lambda_{\text{rank}} = 0.05$
FTID	$k = 32$, $\eta = 1$, $\lambda_{\text{basis}} = 0.001$, $\lambda_{\text{ccorr}} = 0.01$
FPCB	$M = 128$, $\beta_{\text{cb}} = 0.25$, $\lambda_{\text{cb}} = 0.01$, $\lambda_{\text{usage}} = 0.002$, $\lambda_{\text{cb-sub}} = 0.05$
Task objectives	$\lambda_{\text{aux}}^{\text{eeg}} = 0.3$, $\lambda_{\text{aux}}^{\text{eye}} = 0.3$, $\lambda_{\text{acm}} = 0.5$, $\lambda_{\text{tc}} = 0.03$, $\lambda_{\text{dfs}} = 0.2$
Privacy objectives	$\lambda_{\text{adv}} = 0.3$, $\lambda_{\text{id}} = 0.3$, GRL coefficient $\gamma = 0.4$
Additional settings	contrastive temperature $\tau = 0.2$, ACM threshold $\delta = 20$, dominant-feature masking ratio 0.2

Table 5: Component ablation results on SEED-IV and SEED-V. LOSO ACC measures cross-subject task utility, whereas BCA and fused UIA are obtained under leave-one-session-out evaluation. Higher ACC and BCA are better, while lower UIA is better.

	Task utility (%)		**Privacy** (%)
Variant	**LOSO ACC** ↑	**BCA** ↑	**Fused UIA** ↓
SEED-IV			
Full PrivFusionNet	84.23	71.49	**7.57**
w/o PA-LRF	**84.98**	71.47	7.62
w/o FTID	83.91	71.43	7.66
w/o FPCB	84.88	**71.73**	9.77
SEED-V			
Full PrivFusionNet	88.30	**66.35**	**6.11**
w/o PA-LRF	88.10	66.02	7.21
w/o FTID	88.45	65.75	7.56
w/o FPCB	**88.89**	64.63	11.97

B.5 Implementation Details

All neural models are implemented in PyTorch. PrivFusionNet is optimized using AdamW (Loshchilov & Hutter, 2019) for 80 epochs with a batch size of 256, an initial learning rate of 3×10^{-4}, and a weight decay of 10^{-4}. The learning rate follows a cosine annealing schedule. Emotion-classification losses use class-balanced cross-entropy with label smoothing of 0.05. The representation dimension is $d = 128$, and the dropout rate is 0.25.

The main hyperparameters are summarized in Table 4. Detailed definitions of the training objectives are provided in Appendix A.

During training, Gaussian noise with a standard deviation of 0.03 is added to EEG and eye-movement features. Feature, channel, frequency-band, and modality dropout are applied with probability 0.05 and disabled during evaluation.

The identity classifier used for UIA evaluation contains two hidden layers of width 256 and is trained for 40 epochs using AdamW, with a learning rate of 10^{-3} and weight decay of 10^{-4}. The highest observed UIA is reported.

A fixed random seed is used for all experiments.

C Ablation Analysis

We conduct component-wise ablations to examine how PA-LRF, FTID, and FPCB contribute to task utility and identity privacy. Each variant removes one module from the full PrivFusionNet while keeping all other settings unchanged. Table 5 reports cross-subject LOSO ACC together with cross-session BCA and fused-representation UIA.

PA-LRF and FTID regulate the utility–privacy balance. Removing PA-LRF changes fused UIA by only 0.05 percentage points on SEED-IV, but increases it by 1.10 points on SEED-V. Similarly, removing FTID increases UIA by 0.09 and 1.45 points, respectively, while producing only limited changes in task utility. These relatively small isolated effects should be interpreted in the context of the complete model. FPCB has already compressed the fused representation into subject-shared discrete prototypes and reduced UIA close to random-identification levels. This leaves less residual identity information for the upstream risk-aware weighting of PA-LRF or the soft subspace suppression of FTID to remove. Their roles are therefore primarily to maintain a favorable utility–privacy balance rather than to impose an aggressive information bottleneck independently.

FPCB provides the strongest direct identity suppression. Removing FPCB increases fused UIA from 7.57% to 9.77% on SEED-IV and from 6.11% to 11.97% on SEED-V, corresponding to increases of 2.20 and 5.86 percentage points. In contrast, LOSO ACC changes by only 0.65 and 0.59 points. The BCA increases by 0.24 points on SEED-IV but decreases by 1.72 points on SEED-V. These results indicate that the discrete prototype bottleneck removes fine-grained subject-specific variation at a comparatively small task-utility cost, making FPCB the principal source of direct privacy improvement.

Overall, the results support a layered privacy design. PA-LRF reduces identity-sensitive interactions during fusion, FTID attenuates residual identity directions in the continuous fused space, and FPCB performs the strongest final compression through subject-shared prototypes. Because FPCB can mask part of the marginal effect of the two upstream modules, the small single-module differences for PA-LRF and FTID do not imply that they are redundant. Instead, they show that these modules refine the utility–privacy trade-off once strong code-level privacy protection is already present.

D Extended Experiments and Future Directions

D.1 Extension to Cognitive Load Recognition

To evaluate the applicability of PrivFusionNet beyond emotion recognition, we conduct three-class cognitive load recognition on CL-Drive (Angkan et al., 2024), with low, medium, and high load as the target classes. CL-Drive contains multimodal recordings from 21 participants collected during simulated driving. We use its EEG and eye-movement recordings while retaining the core PrivFusionNet architecture.

Data and preprocessing. We use EEG and eye-movement recordings from 21 subjects. The recordings are divided into non-overlapping 10-second windows, resulting in 2,948 valid paired samples. EEG signals from four channels (TP9, AF7, AF8, and TP10) are detrended, band-pass filtered at 0.4–75 Hz, notch filtered at 60 Hz, and processed using Welch spectral estimation. Log-power features from five frequency bands form a 4×5 EEG input. Eye movements are represented by 32 statistical features describing pupil activity, blinks, fixations, and saccades. Normalization statistics are estimated exclusively from each training fold. No additional alignment or target-domain adaptation is applied.

Compared settings. We compare adaptive fusion without privacy protection, unimodal identity suppression (IDS), and the complete PrivFusionNet. The first setting does not employ explicit identity suppression. Unimodal IDS protects the EEG and eye-movement representations independently but introduces no fusion-level protection. PrivFusionNet applies PA-LRF, FTID, FPCB, and fusion-level adversarial learning. Adaptation to CL-Drive only changes the modality-dependent input dimensions and the task and identity output dimensions.

Evaluation protocols. Cross-subject task generalization is evaluated using 21-fold leave-one-subject-out (LOSO) validation. In each fold, one subject is held out for evaluation and the remaining 20 subjects are used for training.

Because CL-Drive does not contain repeated acquisition sessions, privacy is evaluated using leave-one-temporal-block-out splits. Each 180-second driving segment is divided into three consecutive, non-overlapping 60-second blocks. Two blocks are used for training and the remaining block

Table 6: Task utility and identity leakage on CL-Drive. ACC is evaluated under LOSO, whereas BCA and UIA are evaluated using leave-one-temporal-block-out splits. Higher ACC and BCA and lower UIA are better.

	LOSO (%)	Temporal block (%)			
Setting	ACC ↑	BCA ↑	EEG UIA ↓	Eye UIA ↓	Fused UIA ↓
Original leakage	43.38	52.72	11.43	12.81	35.74
Unimodal IDS	44.97	55.53	**7.82**	**6.59**	36.49
PrivFusionNet	**55.77**	**60.23**	39.02	17.65	**13.94**

for evaluation. This preserves the same 21 identities across the training and evaluation partitions. Identity leakage is measured using user identification accuracy (UIA), with a chance level of $1/21 = 4.76\%$.

D.2 Results and Analysis

Table 6 reports cross-subject task performance and representation-level identity leakage.

Cross-task persistence of fusion-level identity re-leakage. Unimodal IDS reduces EEG UIA from 11.43% to 7.82% and eye-movement UIA from 12.81% to 6.59%. Nevertheless, fused UIA remains essentially unchanged, increasing slightly from 35.74% to 36.49%. Together with the results on SEED-IV and SEED-V, this finding reproduces the same fusion-level re-leakage pattern on a different dataset and downstream task. Identity privacy achieved separately in each modality is therefore not necessarily preserved by multimodal fusion, regardless of whether the fused representation is learned for emotion recognition or cognitive load recognition. This cross-task consistency suggests that fusion-level re-leakage is not specific to a particular label space or evaluation dataset, but represents a broader privacy risk in multimodal representation learning.

Cross-task transferability of PrivFusionNet. Without introducing any CL-Drive-specific privacy module, PrivFusionNet reduces fused UIA to 13.94%, corresponding to absolute reductions of 21.80 and 22.55 percentage points relative to the original leakage model and unimodal IDS, respectively. Meanwhile, it achieves the best task performance, improving LOSO ACC from 43.38% to 55.77% and temporal-block BCA from 52.72% to 60.23%. These results show that the privacy–utility benefits of PrivFusionNet transfer from emotion recognition to cognitive load recognition despite changes in the target labels, input dimensions, and data collection setting. They further indicate that PA-LRF, FTID, and FPCB address identity-sensitive structures arising from multimodal fusion rather than patterns specific to emotion recognition.

D.3 Limitations and Future Directions

The current evaluation remains limited to paired EEG and eye-movement inputs. Future work will extend PrivFusionNet toward a general privacy-preserving multimodal framework that supports diverse modality combinations, missing-modality settings, and transfer across downstream tasks.