



Robust multimodal emotion recognition under missing and incomplete data with cross-modal regeneration

Behzad Mahaseni*, Naimul Mefraz Khan

Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto M5B 2R2, Ontario, Canada.

***Correspondence to:** Behzad Mahaseni, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto M5B 2R2, Ontario, Canada. E-mail: behzad.mahaseni@torontomu.ca

Received: March 26, 2026 **Accepted:** July 01, 2026 **Published:** July 02, 2026

Cite this article: Mahaseni B, Khan NM. Robust multimodal emotion recognition under missing and incomplete data with cross-modal regeneration. *Empath Comput.* 2026;2:202610. <https://doi.org/10.70401/ec.2026.0023>

Abstract

Aims: Multimodal emotion recognition (MER) can outperform unimodal approaches by integrating complementary information from multiple sources. However, real-world applications often involve incomplete or missing modalities, reducing the reliability of existing MER models. This study proposes a framework that remains robust under missing-modality conditions while preserving the advantages of multimodal integration.

Methods: To address this challenge, we propose a cross-modal latent regeneration and attention long short-term memory (CMLR-ALSTM) framework. The framework combines pretrained variational autoencoder encoders with residual projection networks, optimized with L2 loss, to achieve stable latent-space alignment across modalities. Regenerated and available latent representations are integrated using a cross-modal attention mechanism and processed by an LSTM to model temporal dependencies and improve multimodal fusion under incomplete data.

Results: The proposed framework was evaluated on three benchmark datasets under complete, partially missing, and completely missing modality scenarios. Experimental results demonstrate that CMLR-ALSTM achieves up to 17.22% improvement under missing-modality conditions while maintaining competitive performance under complete-modality settings. The results also show that the proposed latent regeneration strategy effectively preserves cross-modal relationships and robust latent representations.

Conclusion: The experimental results confirm the effectiveness of the proposed framework, particularly in realistic environments where data availability is inconsistent. By leveraging CMLR to reconstruct missing representations and modelling temporal dependencies through LSTM, the proposed approach provides a more robust and reliable MER framework for practical deployment. The results across different modality combinations demonstrate its ability to generalize across heterogeneous multimodal settings.

Keywords: Multimodal emotion recognition, cross-modal regeneration, variational autoencoder, long short-term memory, electroencephalogram, facial features

1. Introduction

Emotion is a fundamental part of a human's mental state that contributes to the richness and adaptive quality of human mental life, such as interactions between people, decision making, and education^[1,2]. Human emotion analysis and recognition have been active and challenging research topics within the research community for a long time due to their wide range of applications. Rapid advances and model improvements in deep learning and machine learning techniques have enabled researchers to expand the application of emotion recognition to different areas. The primary focus in the early stages of emotion recognition research was on unimodal approaches that relied on behavioural data or physiological signals, treating each modality separately. Although unimodal approaches provide valuable information and have demonstrated promising performance across studies, they often struggle to accurately capture emotional states when relying on a single data source. A single modality may not sufficiently represent individual differences, be sensitive to noise, or be affected by data variability, which may reduce the generalization capability of emotion recognition systems. This makes it challenging to describe an emotional state comprehensively using a single modality^[3,4].

As a result, there has been growing interest in treating emotion recognition as a multimodal problem that integrates information across multiple data sources. Multimodal emotion recognition (MER) refers to the process of identifying human emotions using a



combination of physiological and behavioural signals, such as facial features and physiological measurements. Emotion recognition systems that leverage multiple modalities offer several notable benefits. By integrating complementary information from multiple sources, multimodal approaches can capture richer representations of emotional states and often achieve higher accuracy than unimodal models. They also tend to be more robust and reliable in diverse real-world scenarios, as different modalities can compensate for one another's limitations. Furthermore, multimodal systems reduce ambiguity by jointly analyzing multiple signals, which is particularly beneficial when emotional expressions are subtle, complex, or mixed^[5,6]. In addition, previous studies suggest that multimodal signals capture richer emotional information than unimodal inputs and more closely reflect natural patterns of human expressive behaviour^[7].

However, in real-world scenarios, it is not always assured that data from all modalities will be consistently available due to many factors such as sensor malfunctions, hardware limitations, privacy constraints, environmental interference, and data transmission issues^[8]. In addition, models must understand the complex intramodal and cross-modal interactions to predict emotion. This challenge highlights the importance of developing methods that can effectively handle incomplete or missing modalities while preserving the advantages of multimodal integration. To address the missing modality problem, several early studies in the research community employed multimodal data fusion techniques, including early, late, and intermediate fusion, combined with various imputation techniques during model inference. However, these approaches often degraded the overall performance of the model. Recent multimodal methods have increasingly focused on model based approaches that learn a joint or coordinated representation across modalities^[9-11]. Despite promising results when modalities are partially missing, these methods usually lack robustness under more challenging conditions where an entire modality or more than one modality is missing. In addition, in practical settings, the pattern of missing modalities is often uncertain, where one or more modalities are partially or entirely absent in a random form. To address these challenges, a model should be capable of handling situations where multiple modalities are missing (whether partially or entirely) while still learning robust joint representations under such uncertainty. This work introduces a regenerative network into the variational autoencoder (VAE) to effectively handle missing modalities. To further capture temporal dynamics and classify emotions from sequential data, we incorporate an attention-enhanced long short-term memory (LSTM).

In summary, the main contributions of this work are as follows:

- We propose a cross-modal latent regeneration (CMLR) framework for robust MER under incomplete modality conditions. The proposed framework integrates latent regeneration, attention-based multimodal fusion, and temporal modelling to improve robustness when one or more modalities are partially or completely unavailable during inference.
- We introduce a cross-modal attention-LSTM to model sequential dependencies, enabling adaptive interaction between modalities and improving the quality of multimodal feature integration.
- We demonstrate that the proposed method maintains robust performance and achieves 5.33%-17.48% improvement compared to existing baseline models.
- To further evaluate the generalization capability of the proposed framework, we additionally include experiments using different combinations of modalities beyond electroencephalogram (EEG) and facial data. This allows us to assess whether the proposed latent regeneration and attention-based framework can generalize across heterogeneous multimodal settings, demonstrating its flexibility beyond specific modality pairs.

The remainder of this paper is organized as follows. [Section 2](#) reviews existing studies on MER and discusses challenges associated with incomplete modalities. [Section 3](#) presents the proposed framework for emotion recognition under incomplete data conditions. [Section 4](#) outlines the training procedure. [Section 5](#) describes the datasets, presents the experimental results, and evaluates the performance of the proposed approach. Finally, [Section 6](#) concludes the paper and discusses potential directions for future research.

2. Related Work

Research in psychology has shown that human emotions encompass complex combinations of subjective feelings, physiological reactions, and behavioural responses triggered by internal or external stimuli^[12]. Early research in emotion recognition tasks was primarily focused on uni-modal approaches where the different modalities are treated independently. Researchers have analyzed a wide range of behavioural or physiological signals, such as facial expressions^[13-17], speech^[18-20], gestures^[21], electrocardiogram (ECG)^[22,23], and EEG^[24-26], where each of these approaches has very distinct statistical properties. Among these modalities, EEG signals and facial expressions are widely used for human emotion analysis. As discussed in [Section 1](#), although unimodal approaches have demonstrated promising performance across various studies, they are inherently limited in capturing the full complexity of human emotional states when relying on a single modality. In addition, behavioural data and physiological signals are the primary sources for identifying human emotions. However, behavioural data often have reliability issues and can be ineffective since humans may involuntarily or consciously conceal their real emotions (so-called social masking)^[27]. Moreover, collecting physiological signals relies on specialized equipment during sample acquisition, which is often invasive, expensive, and needs technical guidance from professionals^[12]. In addition, the complexity of emotions and individual differences in physiological responses may lead to declining

prediction performance from a classification perspective when using a single physiological signal^[28]. These limitations have motivated researchers to consider emotion recognition as a multimodal task.

The rest of this section presents an overview of the prior work on MER and a brief overview of related deep learning architectures used in emotion recognition.

2.1 MER

MER aims to infer human emotional states by integrating behavioural indicators and physiological responses obtained from multiple data sources. Compared with unimodal approaches, multimodal frameworks generally provide more reliable predictions because complementary modalities can compensate for limitations present in individual signals. Emotional responses are inherently complex and vary across individuals, which can introduce ambiguity when relying on a single modality. By jointly considering multiple sources of information, multimodal systems can reduce uncertainty and provide a more comprehensive representation of affective states^[29,30]. Several studies have explored MER using diverse combinations of sensory inputs and machine learning techniques. For instance, the studies^[31,32] developed frameworks that integrate facial expression and body gesture information to improve emotion prediction performance. In physiological-based emotion analysis, Dessai and Virani^[33] investigated early fusion of ECG and galvanic skin response (GSR) features and evaluated multiple classifiers, showing improved classification performance when combining modalities. A hypercomplex multimodal neural network was proposed in the work^[34], where parameterized hypercomplex multiplications were used to model interactions between EEG, ECG, GSR, and eye-tracking signals, demonstrating improved modelling of cross-modal relationships. Similarly, Pan *et al.*^[35] introduced a deep learning architecture combining facial expression, speech, and EEG modalities using specialized subnetworks designed for each signal type. Transformer-based approaches have also been investigated, such as the Token-disentangling Mutual Transformer proposed in the work^[36], which separates modality-specific and shared emotional characteristics through token-level representation learning. Additional multimodal architectures integrate physiological and visual signals for emotion classification. For example, Ali and Hughes^[37] proposed a unified transformer-based framework combining facial images with ECG or PPG signals, demonstrating improved performance compared to unimodal configurations. In the study^[38], a dual-branch neural architecture was introduced to jointly process EEG and eye movement signals, where multi-scale fusion and attention mechanisms were applied to improve feature selection. The Transformer Autoencoder presented in the study^[7] addresses missing modality challenges by learning intra- and inter-modal representations using transformer encoders, followed by reconstruction and classification components. Other studies have explored self-supervised representation learning strategies for multimodal emotion analysis^[39,40], leveraging transformer architectures to capture relationships across heterogeneous data sources. A number of works have specifically focused on emotion recognition using EEG signals and facial videos, which are also considered in this study. Several approaches^[41–45] have investigated joint modelling of these modalities using deep learning techniques. Among these, Chen *et al.*^[44] proposed a convolutional network (CNN)-based framework that extracts spatial EEG features using convolutional kernels of different receptive fields and incorporates attention mechanisms to enhance facial feature representations. The work^[45] applied CNN-based facial feature extraction together with bidirectional LSTM modelling of EEG signals, where classification results from each modality were combined using evidence theory-based fusion. Another related study^[46], introduced a decision-level fusion strategy combining EEG signals, peripheral physiological signals, and facial expression data within the valence-arousal framework.

Recent studies have also explored alternative modality combinations such as EEG with eye movement signals. For example, Song *et al.*^[47] proposed a cosine-similarity-based fusion framework that integrates EEG and eye movement representations using gated attention mechanisms, demonstrating strong performance on the SEED-IV dataset. Cross-modal interaction mechanisms have also been investigated to exploit complementary modality information better. In the study^[48], a cross-modal guidance network was introduced where EEG representations are used to refine eye movement features through adaptive weighting and feature recalibration, improving emotion recognition accuracy.

Multimodal learning frameworks offer advantages over unimodal approaches by providing richer and more complementary information, which can improve the reliability and clarity of emotion recognition systems. Accurately predicting emotional states requires the model to capture both relationships within each modality and interactions across modalities. In practical scenarios, multimodal systems must also remain stable when some inputs are noisy, partially corrupted, or unavailable during inference. A key challenge in multimodal learning lies in effectively integrating heterogeneous sources of information into a unified representation^[49]. Multimodal fusion strategies aim to combine features extracted from different data sources, such as visual, physiological, or textual signals, to improve predictive performance. Early research explored several categories of fusion strategies, including early fusion, where modalities are combined at the input level, late fusion, where predictions from individual modalities are aggregated, and intermediate fusion, where interactions are learned at the feature level^[50,51]. Several recent studies have explored robust MER using advanced fusion and cross-modal learning strategies, including GCNet^[52], SDR-GNN^[53], CIF-MMIN^[54], and MoMKE^[55]. These methods utilize mechanisms such as graph neural networks, modality interaction learning, and incomplete modality handling to improve emotion recognition performance under challenging conditions. Moreover, some recent approaches increasingly rely on representation learning techniques that enable the model to discover shared structures across modalities automatically. Among

these approaches, multimodal variational autoencoders (MMVAEs) have emerged as a prominent framework, extending the standard VAE formulation to handle multiple modalities jointly^[56]. MMVAEs are designed to learn both shared and modality-specific latent representations, facilitating cross-modal understanding and improving the effectiveness of multimodal integration. In recent years, a growing body of work has explored MMVAE-based architectures for multimodal representation learning and fusion^[57-61]. In this study, we proposed a MER framework that combines latent regeneration, cross-modal attention, and temporal sequence modelling to improve robustness under incomplete modality scenarios.

3. Proposed Method

The proposed framework for robust MER is composed of three main components: 1) a VAE for modality-specific representation learning, 2) a regenerative joint VAE encoder for cross-modal latent projection and regeneration, and 3) a cross-modal Attention LSTM for sequential emotion classification.

3.1 The VAE

VAE^[62] is a probabilistic generative model that encodes input data into a latent representation and reconstructs it from latent samples. Unlike traditional autoencoders, the VAE constrains latent variables to follow a standard Gaussian prior $N(0, I)$. This constraint ensures that the latent space captures meaningful variations in the data while preserving consistency with the prior distribution. The model is trained by maximizing the evidence lower bound:

$$\mathcal{L}(x; \theta, \phi) = -\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] + D_{\text{KL}}(q_\phi(z|x) \| p(z)) \quad (1)$$

where the reconstruction term ensures the accuracy between the input and output, and the Kullback-Leibler (KL) divergence regularizes the approximate posterior. Therefore, the VAE modules provide probabilistic embeddings for each modality. First, we trained two separate VAEs to capture modality-specific representations of EEG signals and facial images. Each VAE consists of an encoder-decoder pair, generating latent variables z_{eeg} and z_{face} as follows:

$$z_{\text{eeg}} = q_\phi^e(x^{\text{eeg}}), \quad z_{\text{face}} = q_\phi^f(x^{\text{face}}) \quad (2)$$

where q_ϕ^e and q_ϕ^f represent the encoder distributions for the EEG and face modalities, respectively. We adopted a ResNet 101-based VAE pre-trained on diverse external face datasets^[63] for the facial modality, to address limited training data in DEAP and MAHNOB-HCI while enabling effective feature extraction. For the EEG modality, we employed a CNN based encoder, adapted to dataset-specific input representations, consisting of convolutional and max-pooling layers to capture temporal patterns, followed by fully connected layers that estimate the parameters (mean and log-variance) of the latent distribution. This architecture allows the encoder to capture essential temporal dynamics from raw EEG while maintaining a compact and informative latent representation. [Figure 1](#) illustrates the overall architecture of the VAE. Each encoder was trained for 10 epochs with a batch size of 32, using the Adam optimizer with an initial learning rate of 0.001.

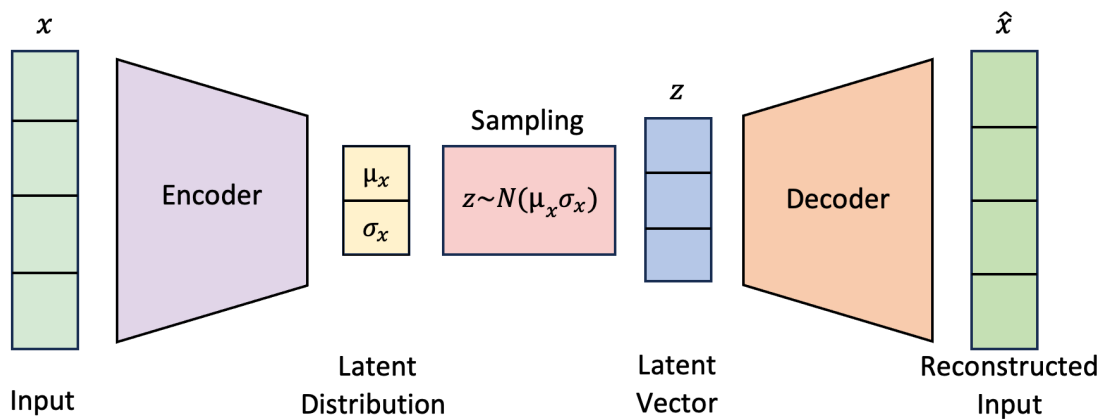


Figure 1. The architecture of VAE. VAE: variational autoencoder.

3.2 CMLR

To address the missing modality problem in MER, we extend unimodal VAEs into a regenerative multimodal framework. Inspired by the study^[64], we introduce cross-modal projections to the pre-trained encoders to regenerate the latent representation of one modality from another. This ensures that informative features can still be exploited even in the absence of other modalities. Given a paired input sample $(x^{\text{eeg}}, x^{\text{face}})$, where x^{face} , we obtain latent codes z_{eeg} and z_{face} from the frozen encoders. Two residual multilayer perceptrons (MLPs) are introduced as projection networks:

$$\hat{z}_{eeg} = G_{f \rightarrow e}(z_{face}), \quad \hat{z}_{face} = G_{e \rightarrow f}(z_{eeg}) \quad (3)$$

where $G_{f \rightarrow e}$ maps from the face latent space to the EEG latent space, and $G_{e \rightarrow f}$ performs the reverse mapping. Both networks share the architecture 128→256→128, ensuring dimensional consistency while enabling non-linear transformations.

3.3 Regeneration loss

To enforce alignment between the original and regenerated representations, we introduce a latent space consistency loss as follows.

$$\mathcal{L}_{reg} = \frac{1}{2} \left(\|\hat{z}_{eeg} - z_{eeg}\|_2^2 + \|\hat{z}_{face} - z_{face}\|_2^2 \right) \quad (4)$$

This regenerative objective ensures that the projected latent vectors remain faithful to their true modality-specific counterparts. By design, this mechanism enables robust inference of missing modalities, since the regenerated latent $\hat{z}$ can substitute for the unavailable z during downstream tasks.

We considered cosine similarity and KL divergence (using the pre-trained decoder) as alternative objectives. Cosine similarity was less effective, as it only enforces directional alignment, while L2 penalizes scale and directional discrepancies. While KL divergence is standard in VAE training, it is not suitable here, since the regeneration networks produce deterministic outputs rather than distributions. Moreover, enforcing Gaussian priors on regenerated latents distorted the modality-specific geometry and degraded downstream classification. For these reasons, we adopt L2-Loss as the regeneration objective.

3.4 Sequential classification with LSTM

After training the cross-modal latent regeneration framework, we used it to extract modality-specific latent representations z_{face} and z_{eeg} from facial and EEG modalities, respectively. Inspired by the work^[65], we implemented a cross-modal temporal attention block to model cross-modal interactions at the sequence level. Formally, the attention mechanism is defined as follows.

$$\text{Attn}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d}} \right) V \quad (5)$$

The mechanism operates with queries $Q = Z_{face}$, keys $K = Z_{eeg}$, and values $V = Z_{eeg}$. This operation yields an attended sequence $\tilde{Z}$, where $\tilde{Z} \in \mathbb{R}^{60 \times d}$ for the DEAP dataset and $\tilde{Z} \in \mathbb{R}^{30 \times d}$ for the MAHNOB-HCI dataset, with d denoting the latent dimensionality, where d is the latent dimensionality. In cases where a modality is missing, the regenerated latent representation $\hat{z}$ from the cross-modal projection is used instead of the unavailable z .

The attended sequence $\tilde{Z}$ was then passed into a two-layer LSTM with 128 hidden units to capture temporal dependencies in the multimodal data. The LSTM selectively retains relevant temporal information through its gating mechanisms and produces a sequence of hidden states $\{h_t\}$. The final hidden state h_T is mapped to emotion predictions via a fully connected layer followed by softmax activation.

$$\hat{y} = \text{softmax}(W \cdot h_T + b) \quad (6)$$

The LSTM classifier was optimized using cross-entropy loss with a dropout rate of 0.3 to reduce overfitting. During training, we considered four combinations of latent inputs: (z_{face}, z_{eeg}) , $(z_{face}, \hat{z}_{eeg})$, $(\hat{z}_{face}, z_{eeg})$, and $(\hat{z}_{face}, \hat{z}_{eeg})$. Each epoch was divided into four passes, one for each pairing, all of which were processed through the attention block and LSTM to update the model with shared weights. The experiments were conducted using 5-fold cross-validation to ensure the reliability and robustness of the results. This architecture enables the model to integrate modality-specific representations from the regenerative framework, capture cross-modal dependencies through attention, and exploit temporal dynamics for robust emotion recognition.

4. Training Procedure

The overall training procedure is composed of three sequential stages: (1) independent pre-training of two VAEs for reconstructing facial images and EEG signals, where the facial branch utilizes a ResNet-101 backbone, (2) optimization of the CMLR module to learn mappings between modality-specific latent spaces, and (3) extraction of latent representations using the trained CMLR framework for subsequent LSTM-based emotion classification. To ensure consistency across stages, the same hyperparameter configuration was adopted throughout the training pipeline. These settings were determined based on commonly used configurations reported in previous studies, along with preliminary empirical evaluation. A batch size of 32 was selected to balance computational efficiency and stable convergence on modern GPU hardware. The learning rate was fixed at 0.001, a standard choice that demonstrated reliable convergence behaviour when combined with the Adam optimizer. The dataset was divided into 80% for training and 20% for validation. The models were trained for 10 epochs, as extending the training duration resulted in only minor performance gains while

increasing computational requirements. This configuration allowed the networks to learn stable reconstructions for both EEG and facial modalities without introducing noticeable overfitting or instability. Since the objective of this stage is to obtain informative latent representations rather than enforce strong disentanglement constraints, the weighting parameter was set to $\beta = 1$, corresponding to the standard VAE formulation. This choice maintains an appropriate balance between reconstruction quality and latent space regularization. Although higher β values ($\beta > 1$) can promote more disentangled latent factors, they typically impose a stronger penalty on reconstruction accuracy, which may negatively affect representation quality for downstream tasks^[66]. The pre-trained VAE components were subsequently incorporated into the multimodal framework's facial encoder and decoder. This initialization improved reconstruction performance compared to training solely on the DEAP and MAHNOB-HCI datasets, resulting in more stable and transferable facial feature representations.

4.1 VAE training

Image encoder: Both the DEAP and MAHNOB-HCI datasets contain a relatively small number of participants, which limits the ability to train a robust VAE directly on these datasets while maintaining strong generalization in facial representation learning. To overcome this limitation, we pre-trained a VAE with a ResNet-101 backbone (Figure 2a) using a larger and more diverse collection of facial image datasets aggregated from multiple publicly available sources^[67-71]. The ResNet-101 architecture was selected due to its proven effectiveness in extracting hierarchical visual features, enabling the VAE to learn discriminative facial representations that capture both structural characteristics and expression related information. This strategy improves the ability of the model to generalize across subjects and experimental conditions.

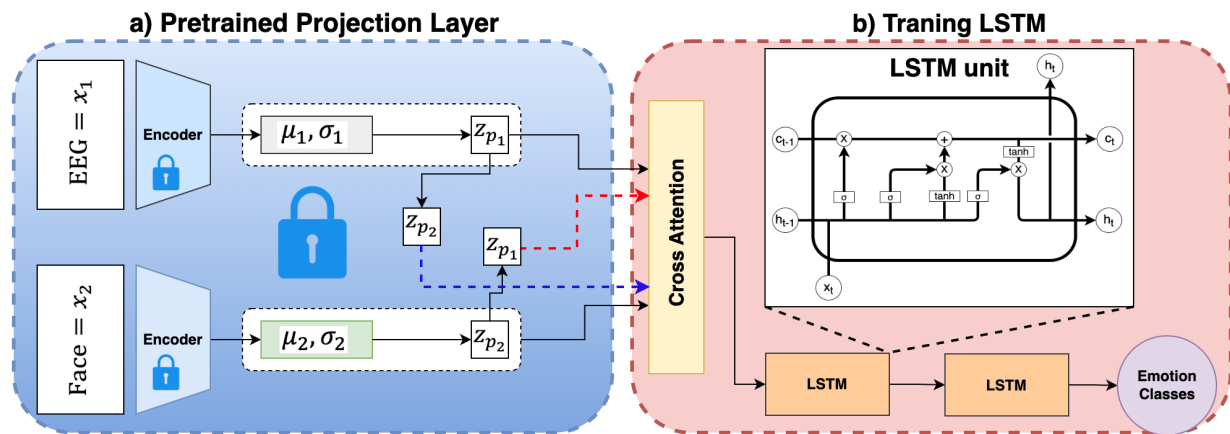


Figure 2. Overview of the proposed model. (a) Two pre-trained encoders are used to obtain actual and regenerated latent representation with the loss computed using regeneration consistency; (b) When one modality is missing (e.g., EEG), the model uses the regenerated latent $\hat{z}_{eeg}$ in place of the unavailable z_{eeg} before passing it through the attention and LSTM modules. EEG: electroencephalogram; LSTM: long short-term memory.

The Chicago Face Database^[67] provides high-quality facial images accompanied by extensive annotations and is widely used in studies related to facial perception and affective computing. The FERET dataset^[68], originally developed to support advances in face recognition research, contains facial images captured under varying illumination conditions, poses, and expressions, making it suitable for learning robust identity-invariant representations. The FEEDTUM dataset^[69] includes video recordings of natural and spontaneous facial expressions, offering valuable temporal variation in emotional displays commonly used for affect recognition research. The MUCT Landmarked Face Database^[70] contains facial images with manually annotated landmark points, facilitating reliable facial alignment and supporting the learning of spatially consistent facial features across individuals. The Person Face Dataset^[71] consists of synthetically generated facial images created using generative adversarial networks, providing additional diversity and enhancing robustness by introducing photorealistic samples. By combining these datasets during pre-training, the facial encoder-decoder benefits from increased variability in identity, pose, and expression, leading to more transferable and stable latent representations. For image-based datasets such as DEAP and MAHNOB-HCI, facial regions were first detected and aligned using multi-task cascaded convolutional networks (MTCNN) to ensure spatial consistency across samples. The extracted facial images were then converted to red-green-blue format, resized to 224×224 pixels, and normalized using dataset-specific statistics before being provided as input to the VAE.

EEG encoder: In addition to the facial branch, we trained a dedicated VAE for the EEG modality to learn compact and informative latent representations prior to multimodal integration. The EEG encoder was implemented using a convolutional architecture. Specifically, the encoder consists of two convolutional layers with 32 and 64 filters, respectively, each followed by a ReLU activation function, and an adaptive average pooling layer. The resulting representation is then flattened and projected through a fully connected layer to produce a 128-dimensional latent vector. This design enables the model to extract stable and discriminative EEG features suitable for downstream emotion recognition.

For the DEAP and MAHNOB-HCI datasets, the EEG inputs were represented as one-second segments with dimensions corresponding to 32 channels and 128 time samples. In contrast, the SEED-IV dataset provides pre-extracted differential entropy (DE) features computed from 62 EEG channels across five standard frequency bands. To ensure temporal alignment between modalities, each trial was segmented into fixed-length chunks consisting of 10 consecutive windows. As a result, each EEG sample was represented as a tensor of size $62 \times 10 \times 5$, where 62 denotes the number of channels, 10 represents the temporal windows within each segment, and 5 corresponds to the frequency bands. This structured representation preserves both temporal and spectral characteristics of the EEG signals.

Similar to the facial branch, the EEG VAE was trained for 10 epochs using a batch size of 32 and the Adam optimizer with a learning rate of 0.001. Training the EEG VAE separately allowed the model to learn a stable and continuous latent space directly from electrophysiological data, providing a robust foundation for multimodal fusion and subsequent temporal modelling.

4.2 CMLR training

For training the CMLR module, EEG and visual modalities were used, with the visual modality corresponding to facial expression information in the DEAP and MAHNOB-HCI datasets. Since the objective is to learn relationships between latent spaces rather than reconstruct the original inputs, the encoders obtained from the previously pre-trained VAEs are utilized to extract modality-specific latent representations.

The EEG modality is processed using a pre-trained CNN-based encoder to capture relevant EEG signals. For both datasets, each EEG segment has dimensions 32×128 , corresponding to one-second windows sampled at 128 Hz. The encoder transforms these signals into compact latent vectors that retain discriminative information related to emotional states.

For the visual modality, a ResNet-101-based encoder pre-trained in the previous stage is employed to obtain facial latent representations. For video-based datasets such as DEAP and MAHNOB-HCI, one representative frame per second is selected to preserve temporal structure while reducing redundancy. Facial regions are first detected and aligned using MTCNN, then cropped and normalized. The processed images are then passed through the encoder to obtain latent vectors that capture identity-invariant structural and expression-related characteristics. After obtaining latent representations from both modalities, denoted as z_{eeg} and z_{face} , cross-modal relationships are learned directly in the latent space through a regeneration mechanism.

Then, instead of enforcing a shared latent space, two MLPs are introduced to map between modality-specific latent spaces, where $G_{e \rightarrow f}(\cdot)$ and $G_{f \rightarrow e}(\cdot)$ denote the cross-modal projection networks.

While the encoder-decoder networks are pretrained to learn meaningful latent representations, the CMLR training stage focuses on aligning latent spaces. During this stage, the encoder parameters are kept fixed and only the projection networks are optimized using a latent regression objective.

$$\mathcal{L}_{\text{CMLR}} = \|\hat{z}_{face} - z_{face}\|_2^2 + \|\hat{z}_{eeg} - z_{eeg}\|_2^2 \quad (7)$$

To evaluate the effectiveness of the selected latent regression loss, we compare the L2 reconstruction loss and cosine similarity. While the L2 loss measures the numerical difference between the projected and original latent representations, cosine similarity evaluates their directional agreement in the latent space. The results are summarized in Table 1.

Table 1. Performance of the cross-modal latent regeneration network.

Metric	Value
L2 Reconstruction Loss	0.063
Cosine Similarity	0.84

As shown in Table 1, the proposed projection network achieved a low L2 reconstruction loss of 0.063 and a high cosine similarity of 0.84. These results indicate that the projected latent representations closely match the original modality specific embeddings in both magnitude and direction. Based on the findings presented in Table 1, the L2 loss was preliminarily selected as the optimization objective for the CMLR module. Its effectiveness is further evaluated in Section 5.4.

The model was trained using a batch size of 32, a learning rate of 0.001, and 10 training epochs. After training the projection networks, the encoders are used to extract temporal latent sequences for each modality. The learned projection networks can generate approximate latent representations for missing modalities, which are then used for emotion classification, enabling robust performance under both complete and missing modality scenarios, as illustrated in Figure 2b.

4.3 CMLR-ALSTM-LSTM-based emotion classification

After training the CMLR framework, we extracted latent representations from the EEG and visual modalities for sequential emotion

classification. For the DEAP and MAHNOB-HCI datasets, when both modalities were completely/partially available, the original latent features z_{face} and z_{eeg} were used. In cases where a modality was missing, its corresponding latent representation was replaced by the regenerated latent feature produced by the cross-modal mapping network. For the SEED dataset, the original latent representations were used without applying cross-modal regeneration. These latent features were then passed to a cross-modal attention block, where the facial latent representation attends to the EEG latent representation in order to improve multimodal feature integration. This operation produces an attended latent sequence $\tilde{z}$, which captures cross-modal interactions while preserving temporal structure.

Emotional responses are inherently dynamic and evolve through temporal variations, which are reflected in both EEG activity and facial expressions. Consequently, modelling sequential dependencies is important for obtaining reliable emotion predictions. The attended latent sequence, denoted as $\tilde{z}$, was provided as input to a two-layer LSTM network with 128 hidden units to capture temporal relationships within the multimodal representations. The LSTM's gating structure enables the network to preserve informative temporal patterns while reducing the influence of transient noise or inconsistencies arising from partial or degraded observations. Model training was performed using the cross-entropy loss function for five epochs. A dropout rate of 0.3 was applied to improve generalization and mitigate overfitting. The last hidden state, h_t , was then fed into a fully connected layer, followed by a softmax activation function to obtain the final emotion prediction. This classification framework enables the model to combine robust latent representations from the VAE encoders, capture cross-modal dependencies through attention, and exploit temporal dynamics through the LSTM, resulting in a more reliable MER system under both complete and incomplete data conditions.

5. Results and Discussion

In this section, we provide a detailed description of the accuracy of emotion classification. Since emotion recognition has a significant impact on many areas, to better understand the classification output, we present emotion classification results in addition to classification robustness.

5.1 Dataset

To evaluate the performance of the proposed MER framework, we conducted experiments on three widely used benchmark datasets: DEAP^[72], MAHNOB-HCI^[73], and SEED IV^[74]. These multimodal human emotion datasets are designated exclusively for academic research, where each participant signed a consent form stating that their data can be used for research and imagery can be published. The multimodal emotion datasets summarized in Table 2 are publicly available for research purposes, and all participants provided informed consent allowing the use of their data for scientific analysis, including the publication of visual materials. The DEAP dataset includes EEG and peripheral physiological recordings from 32 participants, along with frontal facial videos available for 22 of these participants. Each participant viewed 40 one-minute video clips designed to elicit emotional responses. The MAHNOB-HCI dataset provides multimodal recordings from 30 participants, including EEG, facial video, audio, gaze information, and additional physiological signals, with stimulus durations ranging approximately from 35 to 117 seconds.

Table 2. Statistics of the datasets used in this work.

Criterion	DEAP	MAHNOB-HCI	SEED-IV
Dataset size	22 subjects × 40 trials	30 subjects × 20 trials	15 subjects × 3 sessions × 24 trials
Stimulus duration	60 s	49.7-117 s	Varies per film clip
Modality	Vision, EEG, Peripheral	Vision, EEG, Peripheral	EEG, Eye-movement
Used bio signals	EEG, EOG, EMG, GSR, Respiration belt, Plethysmograph, Temperature	EEG, ECG, GSR, Respiration belt, Temperature	EEG (62 channels, 200 Hz), Eye-movement
Labels	Continuous valence and arousal from 1 to 9	Discrete valence and arousal from 1 to 9	Four-class emotion labels: {neutral, happy, sad, fear}

EEG: electroencephalogram; EOG: electrooculography; EMG: electromyography; GSR: galvanic skin response.

To improve data reliability and comparability across subjects, the EEG recordings in both datasets were preprocessed by the dataset providers using standardized procedures. These steps include downsampling the signals to 128 Hz, removal of physiological and environmental artifacts, band-pass filtering within the 4-45 Hz frequency range, re-referencing using a common average reference, and normalization. Emotional annotations in both datasets are based on self-assessment ratings defined on a 1-9 scale according to Russell's valence-arousal model^[75]. For the classification task, ratings between 1 and 5 were considered as low levels, while ratings between 5 and 9 were treated as high levels, resulting in binary labels for both valence and arousal dimensions. In the case of the

DEAP dataset, only participants 1–22 were included in the experiments, as facial video recordings are available only for this subset.

We excluded nine trials where objects/hands covered the participant’s face during part of the trial, and the final data segments consisted of 60 seconds after removing the 3-second pre-trial baseline. For MAHNOB-HCI, 526 valid trials were used, where the central 30 seconds were selected for each trial. The EEG was aligned with facial data by dividing the trials into 1-second segments and extracting the corresponding middle video frame for each EEG segment.

The SEED-IV dataset consists of synchronized EEG recordings and eye-movement signals collected from 15 participants across three experimental sessions, with each session comprising 24 emotion-elicitation trials. In each trial, participants watched a short clip while their EEG and eye-movement signals were recorded. The clips were selected to induce one of four emotional states, including neutral, happy, sad, and fear. EEG data were recorded using a 62-channel acquisition system, downsampled to 200 Hz, and filtered into five conventional frequency bands (delta: 1–4 Hz, theta: 4–8 Hz, alpha: 8–14 Hz, beta: 14–31 Hz, gamma: 31–50 Hz). The dataset providers extracted DE features for EEG and statistical descriptors for eye-movement signals, resulting in 1,080 paired EEG samples with a feature dimension of 62×10 and 1,080 corresponding eye-movement samples with a feature dimension of 32×10 . In this study, we included only trials for which both EEG and eye-movement modalities were fully available, ensuring consistent multimodal input and excluding incomplete or partially missing modality pairs from downstream analysis. Unlike DEAP and MAHNOB-HCI, which include facial video data, the SEED dataset provides eye movement features as the second modality, which are used in place of facial information in our framework.

5.2 Experimental setup

The experiments were performed on a workstation equipped with an NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of VRAM, an Intel Core i9-10920X CPU running at 3.50 GHz, and 64 GB of RAM. The models were implemented in PyTorch 2.1 and trained using CUDA 12.2. All experiments were conducted using an 80%-20% train-validation split, and the reported results correspond to the average performance over 5-fold cross-validation. For all datasets, a subject dependent evaluation strategy was employed. Specifically, all trials from all subjects were first combined and shuffled before splitting. As a result, the same subject could appear in both the training and testing sets, but different trials from that subject were used in each set. Importantly, temporal segments extracted from the same trial were always kept together and were never distributed across both training and testing sets, preventing segment-level data leakage caused by temporally correlated samples. The proposed CMLR-ALSTM framework contains a total of 89,211,844 trainable parameters, including the pretrained encoders, latent projection (regenerative) networks, cross-modal attention module, and LSTM classifier.

For comparison with recent robust MER approaches, we reproduced and adapted the publicly available implementation of the MoMKE^[55] framework for our experimental settings. The adapted implementation was evaluated under fully available modality conditions and complete missing modality scenarios where one modality was entirely unavailable during inference. The original implementation was modified to support the multimodal configurations used in this work, including EEG and face on the DEAP dataset.

5.3 Comparison with baselines and previous work

In this section, the performance of the proposed CMLR-ALSTM framework is evaluated against several baseline approaches and previously reported state-of-the-art methods on the DEAP, MAHNOB-HCI, and SEED-IV datasets. For the DEAP and MAHNOB-HCI datasets, both EEG and facial feature modalities are incorporated, and the corresponding results are shown in Table 3. As shown in Table 3, the proposed method achieves performance comparable to existing approaches on the DEAP dataset while maintaining competitive effectiveness relative to prior studies. Specifically, the proposed method attains 96.93% accuracy for valence and 97.08% for arousal, which is comparable to the best-performing prior models, including the pre-trained disentangled multimodal variational autoencoder (DMMVAE)-LSTM framework. The reproduced MoMKE model achieved performance comparable to our previous DMMVAE-LSTM framework because both approaches rely on learning a joint multimodal latent representation. In particular, both methods combine modality-specific latent embeddings into a shared representation space using a Product-of-Experts-based fusion strategy, enabling the model to capture complementary emotional information across EEG and facial modalities before classification. This indicates that the proposed latent regeneration mechanism can preserve discriminative multimodal features without relying on explicit shared latent-space modelling.

Table 3. Accuracy and Average F1-score for Valence and Arousal on DEAP and HCI (EEG + FACE).

Dataset	Method	Accuracy (Mean $\pm$ SD)	Average F1-score (%)
DEAP	CNN-based ^[44]	Valence: 96.63%	-
		Arousal: 97.15%	-
	LSTM-based ^[45]	Valence: 94.94%	-

	CNN-based ^[46]	Arousal: 95.30%	
		Valence: 77.08%	-
		Arousal: 70.01%	
	LSTM-based ^[63]	Valence: 89.34% $\pm$ 2.8	88.30
		Arousal: 90.67% $\pm$ 2.9	
	Transformer Autoencoder ^[7]	Valence: 96.33%	-
		Arousal: 96.64%	
	MoMKE ^[55]	Valence: 92.46% $\pm$ 2.2	92.23
		Arousal: 93.34% $\pm$ 2.1	
	DMMVAE-LSTM ^[63]	Valence: 92.79% $\pm$ 2.1	92.15
		Arousal: 93.09% $\pm$ 1.9	
HCI	DMMVAE-LSTM with pre-trained VAE ^[63]	Valence: 96.82% $\pm$ 1.8	96.51
		Arousal: 96.90% $\pm$ 1.7	
	AE-ALSTM (with pre-trained AE)	Valence: 95.52% $\pm$ 2.0	95.21
		Arousal: 95.81% $\pm$ 1.8	
	CMLR-LSTM	Valence: 96.11% $\pm$ 1.8	95.96
		Arousal: 96.31% $\pm$ 1.7	
	Ours (CMLR-ALSTM)	Valence: 96.93% $\pm$ 1.8	96.40
		Arousal: 97.08% $\pm$ 1.6	
	CNN-based ^[44]	Valence: 96.69%	-
		Arousal: 96.25%	
	LSTM-based ^[63]	Valence: 92.12% $\pm$ 2.7	91.58
		Arousal: 91.47% $\pm$ 3.0	
	DMMVAE-LSTM ^[63]	Valence: 93.92% $\pm$ 1.8	93.39
		Arousal: 93.18% $\pm$ 1.9	
	DMMVAE-LSTM with pre-trained VAE ^[63]	Valence: 97.12% $\pm$ 1.5	96.88
		Arousal: 96.94% $\pm$ 1.7	
	Ours (CMLR-ALSTM)	Valence: 97.28% $\pm$ 1.5	96.92
		Arousal: 97.05% $\pm$ 1.6	

HCI: human-computer interaction; EEG: electroencephalogram; SD: standard deviation; CNN: convolutional networks; LSTM: long short-term memory; CMLR-ALSTM: cross-modal latent regeneration and attention long short-term memory; VAE: variational autoencoder; DMMVAE: disentangled multimodal variational autoencoder.

For the MAHNOB-HCI dataset, the proposed model consistently outperforms all baseline and prior approaches. In particular, it achieves 97.28% accuracy for valence and 97.05% for arousal, surpassing both the CNN-based and DMMVAE-based methods. This improvement suggests that the cross-modal latent projection strategy effectively captures complementary information between modalities, resulting in more robust and informative feature representations.

To further evaluate the generalization performance of the proposed approach under different combinations of modalities, we also included experiments on the SEED-IV data set, which contains EEG and eye movement signals instead of facial data. The objective of this evaluation is to assess the effectiveness of the proposed latent regeneration and temporal modelling strategy under different multimodal settings. Unlike DEAP and MAHNOB-HCI, the SEED-IV dataset is labeled as a multi-class classification task. Therefore, performance is reported as overall classification accuracy. As shown in Table 4, the proposed CMLR-ALSTM model achieves an accuracy of 93.84%, outperforming all compared methods, including the Transformer Autoencoder (92.69%) and DMMVAE-LSTM (92.77%). The improvement over these approaches demonstrates the effectiveness of the proposed latent regeneration mechanism for learning meaningful cross-modal relationships, even when the modalities differ (facial features vs. eye movement signals).

Table 4. Classification accuracy and F1-score on SEED-IV (EEG + EYE).

Method	Accuracy	F1-score (%)
EECLNet ^[47]	91.16%	-
Cross-Modal Guiding ^[48]	90.21% ± 3.49	-
Transformer Autoencoder ^[7]	92.69%	-
DMMVAE-LSTM with pre-trained VAE	92.77% ± 3.5	92.31
CMLR-ALSTM (Proposed)	93.84% ± 3.2	93.42

EEG: electroencephalogram; EYE: eye movement; LSTM: long short-term memory; DMMVAE: disentangled multimodal variational autoencoder; VAE: variational autoencoder; CMLR-ALSTM: cross-modal latent regeneration and attention long short-term memory.

In addition, we conducted an ablation study using the DEAP dataset to analyze the contribution of the major components of the proposed CMLR-ALSTM framework. Specifically, we evaluated the impact of the attention mechanism and the proposed VAE-based latent representation strategy on emotion recognition performance. We implemented an AE-ALSTM baseline, in which the EEG modality was encoded using a standard autoencoder and facial features were extracted using a convolutional autoencoder. The same attention-based classifier was employed to ensure that the comparison focused on the representation-learning component. To evaluate the contribution of the cross-attention mechanism, we further replaced the cross-attention fusion module in the proposed framework with an intermediate fusion strategy that concatenates the latent representations from both modalities before passing them to the LSTM classifier. This variant is referred to as CMLR-LSTM. The results presented in Table 3 further demonstrate the contribution of the proposed VAE-based latent representation strategy and cross-attention mechanism. Compared with AE-ALSTM, the proposed CMLR-ALSTM framework achieves an average accuracy improvement of 1.34%, highlighting the effectiveness of the VAE-based latent representation in learning a more structured latent space. This structured representation facilitates latent regeneration and enables more informative multimodal embeddings than conventional autoencoders. Furthermore, the proposed CMLR-ALSTM framework achieves an average accuracy improvement of 0.80% over CMLR-LSTM, with the cross-attention module replaced by an intermediate fusion strategy. This result demonstrates the importance of the cross-attention mechanism in capturing complementary information between EEG and facial latent representations. By enabling interactions between modalities before temporal modelling, the attention module produces more discriminative fused representations, thereby improving emotion recognition performance.

The experimental findings across all datasets indicate that the proposed framework consistently outperforms baseline models and previously reported methods. The integration of CMLR with LSTM-based temporal modelling allows the framework to effectively capture both inter-modal relationships and temporal dependencies, contributing to enhanced performance in MER tasks.

5.4 Robustness analysis with missing modality

We further evaluate the robustness of the proposed framework under missing modality conditions, focusing on the extreme case of complete absence (100%) of one modality, as shown in Table 5. To assess this setting, we used our previous models^[63] as baselines and compared their performance under full missing modality scenarios.

Table 5. Valence and Arousal Accuracy and Average F1-score (Mean ± SD) under full missing modality scenarios for DEAP and HCI.

Dataset	Model	Missing	Accuracy (Mean ± SD)	Average F1-score (%)
DEAP	MoMKE ^[55]	Face	Valence: 78.89% ± 3.6 Arousal: 79.29% ± 3.5	78.32
		EEG	Valence: 73.64% ± 3.9 Arousal: 73.77% ± 3.8	73.44
	LSTM	Face	Valence: 68.22% ± 4.5 Arousal: 68.38% ± 4.6	67.88
		EEG	Valence: 63.84% ± 4.7 Arousal: 64.09% ± 4.7	63.43
	DMMVAE-LSTM with pre-trained VAE	Face	Valence: 78.73% ± 2.9 Arousal: 78.12% ± 3.0	77.87
		EEG	Valence: 73.64% ± 3.9 Arousal: 73.77% ± 3.8	73.44

HCI	CMLR-ALSTM without cross modal regeneration	EEG	Valence: 73.71% $\pm$ 3.2 Arousal: 73.85% $\pm$ 3.3	73.16
		Face	Valence: 78.55% $\pm$ 2.9 Arousal: 78.32% $\pm$ 3.0	77.46
		EEG	Valence: 73.92% $\pm$ 3.2 Arousal: 73.53% $\pm$ 3.4	73.08
		Face	Valence: 82.91% $\pm$ 2.2 Arousal: 83.64% $\pm$ 2.3	83.09
		EEG	Valence: 78.96% $\pm$ 2.4 Arousal: 79.18% $\pm$ 2.5	78.72
		Face	Valence: 84.06% $\pm$ 2.0 Arousal: 84.86% $\pm$ 2.1	84.12
	CMLR-ALSTM (Cosine Loss)	EEG	Valence: 80.17% $\pm$ 2.2 Arousal: 80.34% $\pm$ 2.4	79.93
		Face	Valence: 69.89% $\pm$ 4.6 Arousal: 69.94% $\pm$ 4.6	69.47
		EEG	Valence: 64.11% $\pm$ 4.8 Arousal: 64.91% $\pm$ 4.8	64.01
		Face	Valence: 78.99% $\pm$ 3.0 Arousal: 79.45% $\pm$ 2.8	78.79
		EEG	Valence: 74.37% $\pm$ 3.3 Arousal: 74.82 $\pm$ 3.1	74.10
		Face	Valence: 84.97% $\pm$ 2.1 Arousal: 85.38% $\pm$ 2.0	84.81
DMMVAE-LSTM	CMLR-ALSTM	EEG	Valence: 81.07% $\pm$ 2.4 Arousal: 82.39% $\pm$ 2.3	81.32

SD: standard deviation; HCI: human-computer interaction; EEG: electroencephalogram; LSTM: long short-term memory; DMMVAE: disentangled multimodal variational autoencoder; CMLR-ALSTM: cross-modal latent regeneration and attention long short-term memory.

For all baseline methods, the missing modality was approximated by replacing the absent input with the average computed across the dataset. In contrast, the proposed CMLR-ALSTM framework handles missing modalities at the latent level through cross-modal regeneration. Specifically, during inference, when one modality is unavailable, its latent representation (z_{eeg} or z_{face}) is replaced by its regenerated counterpart ($\hat{z}_{\text{eeg}}$ or $\hat{z}_{\text{face}}$), obtained via cross-modal mapping from the available modality. This allows the model to operate effectively under incomplete multimodal inputs.

To further evaluate the contribution of the proposed CMLR mechanism, we included additional experiments in which the regeneration network was not used under the fully missing modality setup. In this setting, when a modality was completely missing, its latent representation was replaced by the average latent representation computed from the training data, as in the baseline methods. In addition, we investigated the effect of the latent regression objective by replacing the proposed L2 reconstruction loss with cosine similarity loss during the training of the cross-modal projection network. This experiment was conducted to assess whether preserving directional alignment alone is sufficient for latent regeneration, compared with directly minimizing the reconstruction error between projected and target latent representations. As shown in Table 5, the CMLR-ALSTM framework without cross-modal regeneration achieves lower performance under fully missing modality conditions than the complete CMLR-ALSTM framework. When the proposed regeneration network is employed, performance consistently improves across all missing-modality scenarios. This observation highlights the contribution of the CMLR mechanism, indicating that the ability to estimate the latent representation of a missing modality from the available modality plays a key role in maintaining robust emotion recognition performance under complete modality absence. Furthermore, the variant trained using cosine similarity loss also achieves lower performance than the proposed L2-based formulation. This result suggests that directly minimizing the reconstruction error between

projected and target latent representations is more effective for CMLR than enforcing directional alignment alone.

Additionally, the results show that CMLR-ALSTM maintains substantially higher performance than the baseline model across all missing modality scenarios for both the DEAP and MAHNOB-HCI datasets. For example, under missing face conditions, the proposed model achieves over 84% accuracy in both valence and arousal on DEAP, compared to approximately 78% for DMMVAE-LSTM and significantly lower performance for the LSTM baseline. A similar trend is observed when the EEG is missing, where the proposed framework consistently outperforms the baselines by a notable margin.

To provide a fair assessment under missing-modality conditions, we also evaluated single-modality LSTM baselines using EEG-only and face-only inputs on the DEAP and MAHNOB-HCI datasets. As shown in the Table 6, EEG-only models consistently outperform face-only models across both datasets, indicating that the EEG modality contributes more strongly to the overall multimodal performance under incomplete modality conditions.

Table 6. Valence and Arousal Accuracy and Average F1-score (Mean $\pm$ SD) for single-modality LSTM experiments on DEAP and HCI datasets.

Dataset	Model	Modality	Accuracy (Mean $\pm$ SD)	Average F1-score (%)
DEAP	LSTM	Face	Valence: 80.18 % $\pm$ 4.2 Arousal: 79.97% $\pm$ 4.5	79.54
		EEG	Valence: 98.77% $\pm$ 0.5 Arousal: 97.98% $\pm$ 0.5	98.17
HCI	LSTM	FACE	Valence: 78.85% $\pm$ 5.2 Arousal: 76.98% $\pm$ 5.5	76.69
		EEG	Valence: 97.65% $\pm$ 0.6 Arousal: 97.45 % $\pm$ 0.5	97.29

SD: standard deviation; LSTM: long short-term memory; EEG: electroencephalogram; HCI: human-computer interaction.

These results demonstrate the effectiveness of the proposed latent regeneration mechanism in preserving discriminative multimodal information even when one modality is completely absent. While joint representation models achieve competitive results when both modalities are available, their performance degrades considerably under full missing modality conditions due to their inability to recover missing information. In contrast, the proposed framework remains significantly more stable by leveraging CMLR, highlighting its practical suitability for real-world multimodal emotion recognition applications where sensor failure or data loss may occur.

5.5 Robustness analysis with data dropout

To evaluate the robustness of the proposed framework under partial data degradation, we conducted additional experiments designed to simulate real-world scenarios, including sensor noise, transmission errors, and temporary signal loss. In this setting, instead of removing an entire modality, we randomly drop a percentage of the input data from both modalities. Specifically, for a given missing rate (e.g., 10%), the dropout is applied equally across modalities (e.g., 5% to EEG and 5% to facial data), ensuring a balanced degradation of multimodal inputs. For this analysis, we used both the DEAP and SEED-IV datasets. These datasets provide synchronized multimodal signals along with temporal sequences, making them suitable for LSTM-based modelling. The temporal structure enables the model to capture sequential dependencies and mitigate the impact of missing or corrupted inputs through temporal aggregation.

The results for the DEAP dataset (Table 7) show that the proposed CMLR-ALSTM framework consistently outperforms DMMVAE-LSTM and the Transformer Autoencoder across all dropout rates. At 10% dropout, the proposed model achieves 93.55% and 93.92% accuracy for valence and arousal, respectively, demonstrating only a minor performance reduction compared to the full-data setting. As the dropout rate increases, degradation remains gradual, indicating that the model effectively utilizes the remaining informative features. Even at 70% dropout, CMLR-ALSTM maintains stable performance (78.37% and 78.94%), outperforming both baseline methods by a clear margin.

Table 7. Valence and Arousal Accuracy and Average F1-score under different data dropout rates (0%-70%) on the DEAP dataset.

Model	Missing Rate	Accuracy (Valence/Arousal)		Average F1-score (%)
DMMVAE-LSTM with pre-trained VAE	0%	Valence: 96.82%	Arousal: 96.90%	96.48
	10%	Valence: 92.67%	Arousal: 92.84%	92.24

CMLR-ALSTM	30%	Valence: 86.22%	Arousal: 86.73%	85.99
	50%	Valence: 79.83%	Arousal: 81.29%	80.03
	70%	Valence: 77.84%	Arousal: 78.32%	77.42
	0%	Valence: 96.93%	Arousal: 97.08%	96.86
	10%	Valence: 93.55%	Arousal: 93.92%	93.30
	30%	Valence: 87.59%	Arousal: 88.02%	87.28
	50%	Valence: 80.69%	Arousal: 81.14%	80.35
Transformer Autoencoder ^[7]	70%	Valence: 78.37%	Arousal: 78.94%	78.08
	0%	Valence: 96.33%	Arousal: 96.64%	-
	10%	Valence: 72.07%	Arousal: 92.33%	-
	30%	Valence: 87.02%	Arousal: 86.03%	-
	50%	Valence: 76.57%	Arousal: 76.45%	-
	70%	Valence: 65.56%	Arousal: 67.01%	-

DMMVAE: disentangled multimodal variational autoencoder; LSTM: long short-term memory; VAE: variational autoencoder; CMLR-ALSTM: cross-modal latent regeneration and attention long short-term memory.

We observed a similar behaviour for the SEED-IV dataset (Table 8), where the proposed framework consistently achieves higher accuracy than competing approaches across all dropout levels. In particular, at higher dropout rates such as 50% and 70%, the performance difference between CMLR-ALSTM and the baseline methods becomes more noticeable, indicating the robustness of the proposed model under more challenging data conditions. This robustness stems from the model's ability to learn complementary representations across modalities, enabling it to leverage the available information when parts of the input are missing. In addition, the LSTM component plays a critical role by integrating temporal information across sequences, smoothing inconsistencies caused by data dropout, and improving classification stability.

Table 8. Classification accuracy and F1-score on the SEED-IV dataset under different data dropout rates (0%-70%).

Model	Missing Rate	Accuracy	F1-score (%)
DMMVAE-LSTM with pre-trained VAE	0%	92.77%	92.31
	10%	84.66%	84.12
	30%	74.18%	73.52
	50%	62.17%	61.34
	70%	48.93 %	47.51
CMLR-ALSTM	0%	93.84%	93.42
	10%	85.81%	85.27
	30%	75.46%	74.93
	50%	63.52%	62.88
	70%	50.17%	48.81
Transformer Autoencoder ^[7]	0%	92.69%	-
	10%	81.76%	-
	30%	67.19%	-
	50%	51.81%	-
	70%	40.00%	-

DMMVAE: disentangled multimodal variational autoencoder; LSTM: long short-term memory; VAE: variational autoencoder; CMLR-ALSTM: cross-modal latent regeneration

and attention long short-term memory.

Overall, these findings demonstrate that the proposed framework is resilient to partial data corruption and maintains stable performance under varying levels of dropout. This makes it well-suited for real-world MER applications, where data quality and availability may fluctuate over time.

5.6 Robustness analysis with modality dropout

An additional experiment was conducted on the DEAP dataset to further examine the robustness of the proposed framework under conditions of partial data loss, reflecting practical challenges such as sensor noise, communication disturbances, or temporary signal degradation. In this setting, instead of fully removing one modality, a certain proportion of the input samples from each modality was randomly discarded during training. This experimental design enables evaluation of the model's capability to operate under incomplete inputs without explicitly reconstructing the missing observations. As presented in Table 9, the CMLR-ALSTM framework maintains stable predictive performance even when a portion of the input data is randomly removed. In comparison with DMMVAE-LSTM, the proposed approach consistently achieves higher classification accuracy across all dropout levels. For example, when 10% of the input data is omitted, only a limited reduction in accuracy is observed, indicating that the model effectively utilizes the remaining informative features. With increasing dropout rates, performance declines gradually rather than suddenly, suggesting that the proposed framework is resilient to partial information loss. This behaviour can be attributed to the combined effect of cross-modal feature interaction and temporal modelling. The interaction between modalities enables the extraction of complementary information, while the LSTM component aggregates temporal context across sequences, reducing inconsistencies introduced by incomplete or noisy inputs. These results indicate that the proposed framework is robust not only in scenarios involving the complete absence of a modality but also when the data are partially corrupted. Maintaining reliable performance under different levels of data degradation highlights the potential applicability of the framework in real-world MER settings, where data quality and availability may fluctuate over time.

Table 9. Accuracy and Average F1-score (Mean $\pm$ SD) under different missing modality scenarios on the DEAP dataset.

Model	Missing rate	Accuracy (Mean $\pm$ SD)	Average F1-score (%)
LSTM	0.0%	Valence: 89.34% $\pm$ 2.8 Arousal: 90.67% $\pm$ 2.9	88.30
	10% Face	Valence: 81.22% $\pm$ 3.1 Arousal: 81.78% $\pm$ 3.0	78.95
	10% EEG	Valence: 78.07% $\pm$ 3.2 Arousal: 78.53% $\pm$ 3.2	75.73
DMMVAE-LSTM with pre-trained VAE	0.0%	Valence: 96.82% $\pm$ 1.8 Arousal: 96.90% $\pm$ 1.7	96.40
	10% Face	Valence: 90.14% $\pm$ 2.0 Arousal: 91.48% $\pm$ 2.1	89.14
	10% EEG	Valence: 87.16% $\pm$ 2.2 Arousal: 87.93% $\pm$ 2.4	86.10
CMLR-ALSTM	0.0%	Valence: 96.93% $\pm$ 1.5 Arousal: 97.08% $\pm$ 1.6	96.92
	10% Face	Valence: 90.34% $\pm$ 2.0 Arousal: 90.91% $\pm$ 2.1	90.20
	10% EEG	Valence: 87.42% $\pm$ 2.2 Arousal: 88.07% $\pm$ 2.4	87.30

SD: standard deviation; DMMVAE: disentangled multimodal variational autoencoder; LSTM: long short-term memory; VAE: variational autoencoder; CMLR-ALSTM: cross-modal latent regeneration and attention long short-term memory; EEG: electroencephalogram.

6. Conclusion

In this study, we proposed a novel MER framework that combines CMLR and attention-enhanced LSTM-based temporal modelling. The proposed CMLR-ALSTM framework learns latent representations of the modality-specific from EEG and visual input. The

cross-modal projection networks regenerate missing latent features when one modality is unavailable. By combining latent-space regeneration with sequential modelling, the framework captures both cross-modal relationships and temporal dependencies inherent to MER.

The experimental results demonstrate that the proposed model achieves competitive, and in several cases superior, performance compared to state-of-the-art unimodal and multimodal approaches across different datasets and experimental settings. On the DEAP and MAHNOB-HCI datasets, the framework achieved strong performance in valence and arousal classification. In contrast, on the SEED-IV dataset, it outperformed previous baseline methods in multi-class emotion recognition using EEG and eye-movement signals. These results confirm the effectiveness of the proposed approach across different modality combinations and benchmark datasets.

Through two robustness-oriented experimental setups, we further validated the stability of the proposed framework. First, under complete missing modality scenarios, the model maintained substantially higher performance than baseline approaches by replacing the unavailable latent representation with its cross-modally regenerated counterpart. Second, under partial data dropout, the proposed approach maintained strong classification performance even at high dropout rates, demonstrating its resilience to gradual data corruption. These findings indicate that the combination of CMLR and temporal learning provides a practical advantage in settings where multimodal inputs may be incomplete or unreliable.

In addition, the findings emphasize the benefits of MER, illustrating how complementary modalities can mitigate limitations of individual data sources and enhance robustness and generalization. The proposed framework's ability to sustain strong performance under both fully available and partially degraded input conditions indicates its suitability for practical applications, where data acquisition may be affected by noise, signal corruption, or limitations in sensing hardware.

Despite the promising results, this study has several limitations. The proposed framework was evaluated on publicly available benchmark datasets using a subject-dependent evaluation protocol. Although this protocol is widely adopted in MER research, evaluating the framework under subject-independent settings and on larger, more diverse datasets would provide a more comprehensive assessment of its generalization to unseen subjects and recording conditions. In addition, the current study focused on EEG, facial images, and eye-movement signals. While the proposed framework demonstrated strong performance across these modality combinations, its applicability to other physiological and behavioural modalities remains to be explored. Future work will include evaluating the proposed framework under subject-independent protocols, investigating more realistic missing-modality scenarios, and validating its performance on larger and more diverse multimodal datasets. In addition, future studies will explore the integration of additional physiological and behavioural modalities, such as ECG and speech signals, to further enhance multimodal representation learning.

Declarations

Authors contribution

Mahaseni B: Methodology, investigation, formal analysis, writing-original draft, writing-review & editing.

Khan NM: Conceptualization, methodology, formal analysis, writing-original draft, writing-review & editing.

Conflicts of interest

The authors declare no conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

Funding

This research was funded by Natural Sciences and Engineering Research Council of Canada (Grant Nos. RGPIN-2020-05471 and Alliance 577563-2022).

Copyright

© The Author(s) 2026.

References

1. Bower GH. Mood and memory. *Am Psychol*. 1981;36(2):129-148. [DOI]
2. Izard CE. Basic emotions, relations among emotions, and emotion-cognition relations. *Psychol Rev*. 1992;99(3):561-565. [DOI][PubMed]
3. Tyng CM, Amin HU, Saad MNM, Malik AS. The influences of emotion on learning and memory. *Front Psychol*. 2017;8:1454. [DOI]
4. LeDoux JE. Emotion circuits in the brain. *Annu Rev Neurosci*. 2000;23(1):155-184. [DOI]
5. Ezzameli K, Mahersia H. Emotion recognition from unimodal to multimodal analysis: A review. *Inf Fusion*. 2023;99:101847. [DOI]
6. Lian H, Lu C, Li S, Zhao Y, Tang C, Zong Y. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*. 2023;25(10):1440. [DOI][PubMed][PMC]
7. Cheng C, Liu W, Fan Z, Feng L, Jia Z. A novel transformer autoencoder for multi-modal emotion recognition with incomplete data. *Neural Netw*. 2024;172:106111. [DOI][PubMed]
8. Wu R, Wang H, Chen HT. Deep multimodal learning with missing modality: A survey. *arXiv:2409.07825v4* [Preprint]. 2024. [DOI]
9. Kumar S, Qiu P, Yang B. Multimodal normative modeling in Alzheimer's disease with introspective variational autoencoders. *BioRxiv* [Preprint] 2024. [DOI]
10. Chen R, Zhou W, Hu H, Fei Z, Fei M, Zhou H. Disentangled variational auto-encoder for multimodal fusion performance analysis in multimodal sentiment analysis. *Knowl Based Syst*. 2024;301:112372. [DOI]
11. Baltrusaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(2):423-443. [DOI]
12. Yang K, Wang C, Gu Y, Sarsenbayeva Z, Tag B, Dingler T, et al. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Trans Affect Comput*. 2023;14(2):1082-1097. [DOI]
13. Li Y, Wang M, Gong M, Lu Y, Liu L. FER-former: Multimodal transformer for facial expression recognition. *IEEE Trans Multimed*. 2025;27:2412-2422. [DOI]
14. Singh R, Saurav S, Kumar T, Saini R, Vohra A, Singh S. Facial expression recognition in videos using hybrid CNN & ConvLSTM. *Int J Inf Technol*. 2023;15(4):1819-1830. [DOI]
15. Sarvakar K, Senkamalavalli R, Raghavendra S, Kumar JS, Manjunath R, Jaiswal S. Facial emotion recognition using convolutional neural networks. *Mater Today Proc*. 2023;80:3560-3564. [DOI]
16. Karatay B, Beştepe D, Sailunaz K, Özyer T, Alhaji R. CNN-Transformer based emotion classification from facial expressions and body gestures. *Multimed Tools Appl*. 2024;83(8):23129-23171. [DOI]
17. Hangaragi S, Singh T, Neelima N. Face detection and recognition using face mesh and deep neural network. *Procedia Comput Sci*. 2023;218:741-749. [DOI]
18. Akinpelu S, Viriri S, Adegun A. An enhanced speech emotion recognition using vision transformer. *Sci Rep*. 2024;14(1):13126. [DOI][PubMed][PMC]
19. Ye J, Wen XC, Wei Y, Xu Y, Liu K, Shan H. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; 2023 Jun 4-10; Rhodes Island, Greece. Piscataway: IEEE; 2023. p. 1-5. [DOI]
20. Kumar S, Haq M, Jain A, Jason C, Moparathi N, Mittal N, et al. Multilayer neural network based speech emotion recognition for smart assistance. *Comput Mater Continua*. 2023;74(1):1523-1540. [DOI]
21. Vijayanthi S, Arunnehr J. Human emotion recognition from body posture with machine learning techniques. In: Singh M, Tyagi V, Gupta PK, Flusser J, Ören T, editors. *Advances in computing and data sciences. Proceedings of the 6th International Conference*; 2022 Apr 22-23; Kurnool, India. Cham: Springer; 2022. p. 231-242. [DOI]
22. Wang L, Hao J, Zhou TH. ECG multi-emotion recognition based on heart rate variability signal features mining. *Sensors*. 2023;23(20):8636. [DOI][PubMed][PMC]
23. Fang A, Pan F, Yu W, Yang L, He P. ECG-based emotion recognition using random convolutional kernel method. *Biomed Signal Process Control*. 2024;91:105907. [DOI]
24. Liu S, Wang Z, An Y, Zhao J, Zhao Y, Zhang YD. EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network. *Knowl Based Syst*. 2023;265:110372. [DOI]
25. Wei Y, Liu Y, Li C, Cheng J, Song R, Chen X. TC-Net: A transformer capsule network for EEG-based emotion recognition. *Comput Biol Med*. 2023;152:106463. [DOI]
26. Iyer A, Das SS, Teotia R, Maheshwari S, Sharma RR. CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings. *Multimed Tools Appl*. 2023;82(4):4883-4896. [DOI]
27. Gunes H, Pantic M. Automatic, dimensional and continuous emotion recognition. *Int J Synth Emot*. 2010;1(1):68-99. [DOI]
28. Zhang J, Yin Z, Chen P, Nichele S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inf Fusion*. 2020;59:103-126. [DOI]
29. Kapase AB, Uke N. A comprehensive review in affective computing: An exploration of artificial intelligence in unimodal and multimodal

- emotion recognition systems. *Int J Speech Technol.* 2025;28(2):541-563. [DOI]
30. Moon AS, Kim H, Park YC, Lee J. A survey on multimodal emotion recognition: Methods, datasets, and future directions. *Comput Mater Contin.* 2026;87(2):1. [DOI]
31. Tseng HT, Hsieh CC, Xu CH. A two-stage multimodal emotion analysis using body actions and facial features. *Signal Image Video Process.* 2025;19(4):313. [DOI]
32. Liu J, Wang Z, Nie W, Zeng J, Zhou B, Deng J, et al. Multimodal emotion recognition for children with autism spectrum disorder in social interaction. *Int J Hum.* 2024;40(8):1921-1930. [DOI]
33. Dessai A, Virani H. Multimodal and multidomain feature fusion for emotion classification based on electrocardiogram and galvanic skin response signals. *Sci.* 2024;6(1):10. [DOI]
34. Lopez E, Chiarantano E, Grassucci E, Comminiello D. Hyper complex multimodal emotion recognition from EEG and peripheral physiological signals. In: 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW); 2023 Jun 4-10; Rhodes Island, Greece. Piscataway: IEEE; 2023. p. 1-5. [DOI]
35. Pan J, Fang W, Zhang Z, Chen B, Zhang Z, Wang S. Multimodal emotion recognition based on facial expressions, speech, and EEG. *IEEE Open J Eng Med Biol.* 2024;5:396-403. [DOI]
36. Yin G, Liu Y, Liu T, Zhang H, Fang F, Tang C, et al. Token-disentangling Mutual Transformer for multimodal emotion recognition. *Eng Appl Artif Intell.* 2024;133:108348. [DOI]
37. Ali K, Hughes CE. A unified transformer-based network for multimodal emotion recognition. *arXiv:2308.14160v1 [Preprint].* 2023. [DOI]
38. Fu B, Gu C, Fu M, Xiao Y, Liu Y. A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals. *Front Neurosci.* 2023;17:1234162. [DOI]
39. Wu Y, Daoudi M, Amad A. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. *IEEE Trans Affect Comput.* 2024;15(1):157-172. [DOI]
40. Wang R, Zhu J, Wang S, Wang T, Huang J, Zhu X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. *Int J Multimed Inf Retr.* 2024;13(4):39. [DOI]
41. Singh P, Tripathi MK, Patil MB, Shivendra , Neelakantappa M. Multimodal emotion recognition model via hybrid model with improved feature level fusion on facial and EEG feature set. *Multimed Tools Appl.* 2025;84(1):1-36. [DOI]
42. Mutawa AM, Hassouneh A. Multimodal real-time patient emotion recognition system using facial expressions and brain EEG signals based on machine learning and log-sync methods. *Biomed Signal Process Control.* 2024;91:105942. [DOI]
43. Chen Y, Bai Z, Cheng M, Liu Y, Zhao X, Song Y. Multimodal emotion recognition for hearing-impaired subjects by fusing EEG signals and facial expressions. In: 2023 42nd Chinese Control Conference (CCC); 2023 Jul 24-26; Tianjin, China. Piscataway: IEEE; 2023. p. 1-6. [DOI]
44. Wang S, Qu J, Zhang Y, Zhang Y. Multimodal emotion recognition from EEG signals and facial expressions. *IEEE Access.* 2023;11:33061-33068. [DOI]
45. Wu Y, Li J. Multi-modal emotion identification fusing facial expression and EEG. *Multimed Tools Appl.* 2023;82(7):10901-10919. [DOI]
46. Zhu Q, Lu G, Yan J. Valence-arousal model based emotion recognition using EEG, peripheral physiological signals and facial expression. In: *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*; 2026 Jan 24-25; Copenhagen, Denmark. New York: Association for Computing Machinery; 2020. p. 81-85. [DOI]
47. Song Y, Feng L, Zhang W, Song X, Cheng M. Multimodal emotion recognition based on the fusion of EEG signals and eye movement data. In: 2024 IEEE 25th China Conference on System Simulation Technology and its Application (CCSSTA); 2024 Jul 21-23; Tianjin, China. Piscataway: IEEE; 2024. p. 127-132. [DOI]
48. Fu B, Chu W, Gu C, Liu Y. Cross-modal guiding neural network for multimodal emotion recognition from EEG and eye movement signals. *IEEE J Biomed Health Inform.* 2024;28(10):5865-5876. [DOI][PubMed]
49. Koromilas P, Giannakopoulos T. Deep multimodal emotion recognition on human speech: A review. *Appl Sci.* 2021;11(17):7962. [DOI]
50. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf Fusion.* 2017;37:98-125. [DOI]
51. Ramachandram D, Taylor GW. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process Mag.* 2017;34(6):96-108. [DOI]
52. Lian Z, Chen L, Sun L, Liu B, Tao J. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Trans Pattern Anal Mach Intell.* 2023;45(7):8419-8432. [DOI]
53. Fu F, Ai W, Yang F, Shou Y, Meng T, Li K. SDR-GNN: Spectral domain reconstruction graph neural network for incomplete multimodal learning in conversational emotion recognition. *Knowl Based Syst.* 2025;309:112825. [DOI]
54. Liu R, Zuo H, Lian Z, Schuller BW, Li H. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Trans Affect Comput.* 2024;15(4):1856-1873. [DOI]
55. Xu W, Jiang H, Liang X. Leveraging knowledge of modality experts for incomplete multimodal learning. In: *Proceedings of the 32nd ACM International Conference on Multimedia*; 2024 Oct 28-Nov 1; Melbourne, Australia. New York: ACM; 2024. p.438-446. [DOI]
56. Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv:1312.6114 [Preprint].* 2022. [DOI]
57. Aguila AL, Chapman J, Altmann A. Multi-modal variational autoencoders for Normative modelling across multiple imaging modalities. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2023 Oct 8-12; Vancouver, Canada. Cham: Springer; 2023. p. 425-434. [DOI]
58. Martí-Juan G, Lorenzi M, Piella G. MC-RVAE: Multi-channel recurrent variational autoencoder for multimodal Alzheimer's disease

- progression modelling. *Neuroimage*. 2023;268:119892. [DOI][PubMed]
59. Shi T, Wei Y, and Kender JR. An efficient and explanatory image and text clustering system with multimodal autoencoder architecture. *arXiv:2408.07791v1* [Preprint]. 2024. [DOI]
 60. Suzuki M, Matsuo Y. A survey of multimodal deep generative models. *Adv Robot*. 2022;36(5-6):261-278. [DOI]
 61. Kutuzova S, Krause O, McCloskey D. Multimodal variational autoencoders for semi-supervised learning: In defense of product-of-experts. *arXiv:2101.07240v2* [Preprint]. 2021. [DOI]
 62. Kingma DP, Welling M. An introduction to variational autoencoders. *Found Trends Mach Learn*. 2019;12(4):307-392. [DOI]
 63. Mahaseni B, Khan NM. Multimodal emotion recognition with disentangled representations: Private-shared multimodal variational autoencoder and long short-term memory framework. *Empath Comput*. 2025;1(2):202507. [DOI]
 64. Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabat M, et al. Self-supervised learning from images with a joint-embedding predictive architecture. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17-24; Vancouver, Canada. Piscataway: IEEE; 2023. p. 15619-15629. [DOI]
 65. Wei X, Zhang T, Li Y, Zhang Y, Wu F. Multi-modality cross attention network for image and sentence matching. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13-19; Seattle, USA. Piscataway: IEEE; 2020. p. 10938-10947. [DOI]
 66. Burgess CP, Higgins I, Pal A, Matthey L, Watters N, Desjardins G, et al. Understanding disentangling in β -vae. *arXiv:1804.03599v1* [Preprint]. 2018. [DOI]
 67. Ma DS, Correll J, Wittenbrink B. The Chicago face database: A free stimulus set of faces and norming data. *Behav Res Methods*. 2015;47(4):1122-1135. [DOI][PubMed]
 68. Phillips PJ, Wechsler H, Huang J, Rauss PJ. The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis Comput*. 1998;16(5):295-306. [DOI]
 69. Wallhoff F, Schuller B, Hawellek M, Rigoll G. Efficient recognition of authentic dynamic facial expressions on the feedtun database. In: 2006 IEEE International Conference on Multimedia and Expo; 2006 Jul 9-12; Toronto, Canada. Piscataway: IEEE; 2006. p. 493-496. [DOI]
 70. Milborrow S, Morkel J, Nicolls F. The MUCT face database [Internet]. Pretoria: Pattern Recognition Association of South Africa; 2010. Available from: <http://www.milbo.org/muct/>
 71. AlmightyJ. Person face dataset (thispersondoesnotexist) [Internet]. San Francisco: Kaggle; 2024. Available from: <https://www.kaggle.com/datasets/almightyj/person-face-dataset-thispersondoesnotexist>
 72. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, et al. DEAP: A database for emotion analysis; using physiological signals. *IEEE Trans Affect Comput*. 2012;3(1):18-31. [DOI]
 73. Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. *IEEE Trans Affect Comput*. 2012;3(1):42-55. [DOI]
 74. Zheng WL, Liu W, Lu Y, Lu BL, Cichocki A. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Trans Cybern*. 2019;49(3):1110-1122. [DOI][PubMed]
 75. Russell JA. A circumplex model of affect. *J Pers Soc Psychol*. 1980;39(6):1161-1178. [DOI]

Publisher's Note

Science Exploration remains a neutral stance on jurisdictional claims in published maps and institutional affiliations. The views expressed in this article are solely those of the author(s) and do not reflect the opinions of the Editors or the publisher.