A UNIFIED FRAMEWORK FOR EEG–VIDEO EMOTION RECOGNITION WITH BRAIN ANATOMY AND TEMPORAL PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies in video- and EEG-based emotion recognition have shown notable progress. However, multi-modal emotion recognition remains largely unexplored, particularly the integration of physiological signals with video. This integration is crucial, as EEG–video fusion combines observable behavioral cues with internal neural dynamics and enables a more comprehensive and robust characterization of human emotion. To this end, we propose EVER, a novel EEG–Video Emotion Recognition framework that effectively integrates complementary information from both modalities. Specifically, EVER models the anatomical organization of multi-channel EEG signals and the temporal relationships among video clips through graph-based representations. These structured features are subsequently integrated with global modality representation using an inter-modal graph for emotion prediction. To provide a comprehensive evaluation, we establish a benchmark across two public EEG-video paired datasets, Emognition and MDMER. We evaluate 12 representative models, consisting of 5 EEG-only, 5 video-only, and 2 EEG-video models. Extensive experiments demonstrate that the proposed EVER achieves state-of-the-art performance by jointly modeling behavioral cues from video and physiological responses, thereby enabling the recognition of emotional patterns unattainable by either modality alone.

1 INTRODUCTION

Emotion recognition is a fundamental task in numerous real-world applications, including mental health monitoring (Aina et al., 2024), human-computer interaction (Chowdary et al., 2023), and affective computing (Cortiñas-Lorenzo & Lacey, 2023). Recent advances leveraging video (Zhang et al., 2021; Ben et al., 2021) or EEG signals (Peng et al., 2022; Liu et al., 2024) have demonstrated the potential of each modality to provide valuable insights into human emotion. Video captures observable behavioral cues such as facial expressions (Zhang et al., 2020), whereas EEG reflects neural dynamics associated with internal states (Lee et al., 2024). Despite their complementary characteristics, these modalities have typically been studied in isolation, and most existing datasets and benchmarks (Koelstra et al., 2011; Zheng & Lu, 2015; Zafeiriou et al., 2017; Katsigiannis & Ramzan, 2017) are designed for unimodal learning.

However, relying on a single modality provides only a partial and potentially unreliable view of human emotion. Video-based methods can be hindered when individuals mask their emotions (Iwasaki & Noguchi, 2016) or when visual cues are degraded by occlusion or adverse lighting (Elsheikh et al., 2024). Conversely, EEG provides access to internal neural states but remains sensitive to inter-subject variability and sensor artifacts (Peng et al., 2022). These complementary properties motivate their joint modeling. Nevertheless, only a few studies have directly explored EEG-video emotion recognition (Lee et al., 2024; Wang et al., 2026), and they mainly fuse globally aggregated features or predictions without explicitly modeling EEG anatomy and video temporal structure. Although recent EEG foundation models provide strong global representation, three challenges remain. First, global pooling discards channel-level anatomical organization. Second, video is often represented by a single aggregated feature, obscuring the temporal evolution of facial behavior. Third, existing EEG-video fusion methods lack a unified mechanism for relating these fine-grained structures to pretrained global representations.

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.

To address these challenges, we introduce a novel EEG-Video Emotion Recognition (EVER) framework that combines strong pretrained backbones with graph-based structural modeling and inter-modal emotion reasoning. For EEG, we introduce a Brain Anatomy-aware Regional Graph Convolutional Network (BAR-GCN). BAR-GCN groups EEG channels into anatomically meaningful regions and models their relationships to construct region-level representations. For video, we introduce a Video Clip Gaussian temporal prior (VCG), which retains individual clip features as temporally ordered nodes. Sinusoidal positional encoding provides clip-order information, while learnable Gaussian weighting regulates clip-to-clip affinities according to temporal distance. Finally, EVER constructs an inter-modal graph over the global EEG embedding, montage-flexible EEG region nodes, and temporally ordered video clip nodes. Unlike conventional EEG-video methods that operate on globally aggregated modality representations, EVER jointly preserves and refines global and fine-grained structural information within a unified graph.

To systematically evaluate our framework, we establish an extensive benchmark on two public datasets including Emognition (Saganowski et al., 2022) and MDMER (Yang et al., 2024). Both datasets provide time-synchronized EEG and video recordings, enabling the joint analysis of observable behavioral cues and neural responses. We enforce subject-independent splits to prevent participant overlap between training and evaluation sets and to evaluate performance on unseen subjects. Our benchmark covers unimodal and multi-modal settings with 12 representative models, including five general video transformers, five EEG-specific foundation and transformer models, and two EEG-video fusion architectures. Given the limited scale of paired EEG-video datasets, we employ pretrained modality-specific backbones and focus on structural modeling and inter-modal fusion. Extensive experiments show that EVER achieves the highest average accuracy among the evaluated methods on both datasets.

2 Related Work

2.1 Video and EEG Emotion Recognition

Video- and EEG-based emotion recognition have been widely studied as unimodal tasks. Existing video approaches often employ task-specific architectures tailored to facial expressions or affective behavior (Zhang et al., 2019; 2021; Ben et al., 2021), while EEG methods typically design specialized sequence or graph networks to capture temporal dependencies and inter-channel relationships (Gao et al., 2022; Du et al., 2022; Pan et al., 2023). Although effective within individual datasets, these models are commonly optimized for individual datasets and often provide limited transferability across recording conditions or acquisition setups.

These limitations motivate the use of pretrained foundation backbones. Large-scale pretraining provides more transferable representations and is particularly valuable for EEG-video emotion recognition, where paired multi-modal datasets are relatively small. Accordingly, we benchmark representative video transformers, including ViViT (Arnab et al., 2021), TimeSformer (Bertasius et al., 2021), Video Swin Transformer (Liu et al., 2021), VideoMAE (Tong et al., 2022), and AdaMAE (Bandara et al., 2023). We also evaluate EEG-specific foundation models, including BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), and REVE (El Ouahidi et al., 2025), together with ST-Transformer (Song et al., 2021). These models provide strong modality-specific representations and enable a more reliable comparison of unimodal performance. Despite advances in pretrained representation learning, EEG-video fusion remains underexplored. Existing methods generally combine globally pooled modality features or aggregate instance-level predictions (Lee et al., 2024; Wang et al., 2026). These methods neglect the anatomical organization of EEG channels and the temporal evolution of facial behavior.

To address these limitations, EVER augments general pretrained backbones with structured graph reasoning. BAR-GCN constructs anatomically informed regional EEG representations, the Video Clip Gaussian temporal prior (VCG) regulates relationships among temporally ordered video clips, and an inter-modal graph integrates these structured nodes with the global EEG embedding.

2.2 Graph Convolutional Networks

Graph Convolutional Networks (GCNs) (Kipf & Welling, 2017) have been widely adopted in EEG emotion recognition to model dependencies among electrodes. Existing methods typically construct

graphs based on spatial distance, functional connectivity, or anatomical priors (Gao et al., 2022; Du et al., 2022; Pan et al., 2023). Several studies further model EEG signals at multiple spatial scales. HD-GCN (Ye et al., 2022) captures local and global dependencies through separate graph branches. PGCN (Jin et al., 2024) aggregates electrode features at local, regional, and global levels, while MS-GCN (Du et al., 2022) models interactions among channels and predefined regions. Although these methods exploit regional EEG structure, their graph definitions are often tied to the electrode layout, spatial coordinates, or regional partition of a specific montage. Applying them to datasets with different channel configurations therefore requires reconstructing the graph or redefining its regional organization. In addition, their reasoning remains confined to EEG and does not explicitly connect regional neural activity with another modality.

Graph reasoning has also been explored in video understanding by treating objects or temporal segments as nodes (Wang & Gupta, 2018; Yan et al., 2018; Yang et al., 2020). These methods commonly rely on appearance similarity, local temporal connections, or predefined relations. Learnable Gaussian kernels have also been introduced for long-range temporal modeling (Piergiovanni & Ryoo, 2019), but primarily in the form of temporal convolution rather than graph interaction among pretrained clip features. Consequently, existing EEG and video graph methods remain modality-specific and do not provide a unified structure for multi-modal emotion reasoning.

To address these limitations, EVER introduces montage-flexible anatomical modeling and graph reasoning over pretrained modality representations. BAR-GCN maps electrodes from heterogeneous montages into five canonical scalp regions defined by the international 10-20 system (Herwig et al., 2003; Macorig et al., 2021), producing region nodes without dataset-specific graph reconstruction. The VCG temporal prior converts pretrained video features into temporally ordered clip nodes and modulates their affinities using positional encoding and learnable Gaussian distance weights. These fine-grained EEG and video nodes are jointly refined with the global EEG foundation-model representation through a unified inter-modal graph. Unlike prior methods that either perform modality-specific graph reasoning or fuse globally pooled features, EVER directly connects pretrained global representations with anatomy-aware EEG regions and temporally structured video clips.

3 METHOD

3.1 OVERALL FRAMEWORK

Figure 1 illustrates our EVER framework. In detail, it consists of the following three components.

3.1.1 VIDEO NETWORK

We adopt MAE-based video backbones (Tong et al., 2022; Bandara et al., 2023) to extract visual representations. Given a video $\mathbf{V}$, we divide its full temporal sequence into N ordered clips $\{\mathbf{V}_n\}_{n=1}^{N}$, where N varies with the video duration. Each clip is uniformly sampled into 32 frames and represented as $\mathbf{V}_n \in \mathbb{R}^{32\times3\times H\times W}$, where H and W denote the frame height and width. We retain the pretrained encoder f_v and discard its decoder. All encoder blocks are frozen except the final block, which is fine-tuned jointly with EVER. The shared encoder extracts each clip representation as:

$$\mathbf{g}_n = f_v(\mathbf{V}_n) \in \mathbb{R}^d, \tag{1}$$

where $d = 768$ is the embedding dimension. This yields variable-length ordered clip embeddings.

3.1.2 EEG NETWORK

Let the EEG input paired with the corresponding video be $\mathbf{E} \in \mathbb{R}^{C\times T}$, where C and T denote the number of channels and temporal samples, respectively. The complete EEG sequence is resampled to 2,000 samples, globally z-score normalized, and divided into ten 200-sample patches for the pretrained CBraMod (Wang et al., 2025) encoder as follows:

$$\mathbf{z}_e = f_e(\mathbf{E}) \in \mathbb{R}^d. \tag{2}$$

In parallel, the same raw EEG signal is processed through a separate channel-wise projection pathway for BAR-GCN. This dual-path design captures global temporal dynamics through the pretrained encoder while preserving channel-level anatomical structure through BAR-GCN. Since each label describes the complete stimulus interval rather than individual clips, we use a global EEG node while preserving local temporal variation through the video clip nodes.

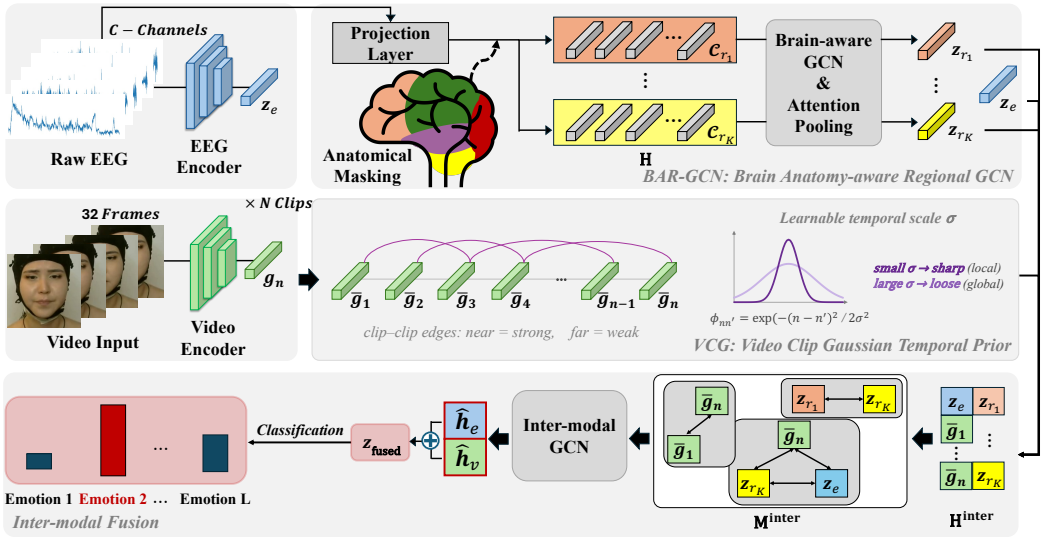


Figure 1: **Overall framework of the EEG-Video Emotion Recognition (EVER).** The global EEG embedding $\mathbf{z}_e$ and the temporally ordered video clip embeddings $\bar{\mathbf{g}}_n$ are extracted from synchronized raw EEG and video. In BAR-GCN, the per-channel EEG projection yields local nodes $\mathbf{H}$, which perform BAR-GCN with anatomical masking within region $\mathcal{C}_{r_k}$ and attention pooling to produce region features $\mathbf{z}_{r_k}$. The VCG temporal prior represents the video as N temporally ordered clip nodes $\{\bar{\mathbf{g}}_n\}_{n=1}^{N}$ and modulates their affinities using the learnable Gaussian weight $\phi_{nn'}$. In the inter-modal fusion stage, we form $\mathbf{H}^{\text{inter}}$ and apply a mask $\mathbf{M}^{\text{inter}}$. Finally, emotion is predicted from the fused logit $\mathbf{z}_{\text{fused}}$ derived from $\hat{\mathbf{h}}_e$ and $\hat{\mathbf{h}}_v$.

3.1.3 Fusion Module

Given the unimodal representations, EVER performs fusion through a unified inter-modal graph. BAR-GCN organizes channel-level EEG features into anatomically grounded region nodes, while the VCG temporal prior equips the video clip nodes with explicit temporal order and distance information. These structured nodes, together with the global EEG embedding, form the inter-modal graph. GCN propagates information across the two modalities, and the refined EEG and video node representations are used to produce the emotion logits.

3.2 Brain Anatomy-aware Regional GCN

The raw EEG signal $\mathbf{E} \in \mathbb{R}^{C\times T}$ is processed independently of the global foundation-model pathway. For each channel, we compute the log-magnitude STFT using a Hann window, with $n_{\text{fft}} = 512$ for Emognition and 128 for MDMER and a hop size of $n_{\text{fft}}/4$. The resulting spectrum $\tilde{\mathbf{E}}$ is resized to 128×256, flattened, and projected to dimension d as:

$$\mathbf{H} = \text{proj}\left(\text{flatten}(\tilde{\mathbf{E}})\right) \in \mathbb{R}^{C\times d}, \tag{3}$$

where $\mathbf{H}$ is the channel-level node matrix and $\mathbf{h}_i \in \mathbb{R}^d$ denotes the embedding of the i-th channel. This projection preserves the full frequency-time input of each channel without temporal pooling, while compressing it into the common graph dimension d.

Motivated by evidence that brain anatomy emphasizes region-specific interactions, we design a region-level graph that embeds anatomical priors and enables the network to capture intra-region relationships through GCN. Let $\mathcal{R} = \{r_k\}_{k=1}^{K}$ denote K predefined brain regions. Each region r_k is associated with a subset of EEG channels, represented by the index set $\mathcal{C}_{r_k} \subseteq \{1,\dots,C\}$. Based on this anatomical prior, the local adjacency matrix $\tilde{\mathbf{A}}_{ij}$ is computed by cosine similarity with region-wise masking as follows:

$$\tilde{\mathbf{A}}_{ij} = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\|\,\|\mathbf{h}_j\|} \cdot \mathbf{M}_{ij}^{\text{local}}, \quad i,j \in [1,C]. \mathbf{M}_{ij}^{\text{local}} = \begin{cases} 1, & \text{if } \exists k \text{ s.t. } i,j \in \mathcal{C}_{r_k} \text{ and } i \neq j, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathbf{M}^{\text{local}} \in \{0,1\}^{C\times C}$ encodes whether channels i and j belong to the same brain region r_k. We retain signed cosine similarities without rectification, allowing dissimilar nodes to exert subtractive influence. After masking and symmetrization, the degree is computed from the signed row sum and lower-bounded by $\epsilon = 10^{-3}$ before normalization. We first obtain the symmetrically normalized adjacency matrix and then apply a two-layer GCN over all EEG channels as follows:

$$\hat{\mathbf{A}} = \mathbf{D}^{-1/2}\tilde{\mathbf{A}}\mathbf{D}^{-1/2}, \;\; \mathbf{H}^{\text{local}} = \text{GCN}*2\left(\text{GELU}\left(\text{GCN}*1(\mathbf{H}, \hat{\mathbf{A}})\right), \hat{\mathbf{A}}\right), \tag{5}$$

where $\mathbf{D}$ denotes the degree matrix of the symmetrized adjacency $\tilde{\mathbf{A}}$, and $\hat{\mathbf{A}}$ is its symmetrically normalized form. $\mathbf{H}$ denotes the input channel representations, while $\mathbf{H}^{\text{local}}$ represents the region-aware channel features obtained by two GCN layers (Kipf & Welling, 2017) with a GELU activation between them.

To obtain compact region-level representations, we group the channel outputs according to the anatomical partition $\mathcal{R}$ and apply attention pooling within each brain region r_k. Specifically, attention pooling assigns a learnable weight $\alpha_{k,i}$ to each channel representation $\mathbf{h}_i^{\text{local}}$ within a region using a parameter vector $\mathbf{w} * a$, and aggregates them into a region-level embedding $\mathbf{z} * r_k \in \mathbb{R}^d$ as:

$$\alpha_{k,i} = \frac{\exp\left(\mathbf{w}_a^\top \mathbf{h}_i^{\text{local}}\right)}{\sum_{j\in\mathcal{C}_{r_k}} \exp\left(\mathbf{w}_a^\top \mathbf{h}_j^{\text{local}}\right)}, \;\; \mathbf{z}_{r_k} = \sum_{i\in\mathcal{C}_{r_k}} \alpha_{k,i}\mathbf{h}_i^{\text{local}}. \tag{6}$$

3.3 Video Clip Gaussian Temporal Prior

While the EEG branch models fine-grained channel interactions guided by anatomical priors, encoding the whole video into a single embedding collapses the entire sequence into one vector and discards the temporal evolution of facial behavior, which is a salient cue for emotion.

To preserve this temporal structure, we introduce a Video Clip Gaussian temporal prior (VCG). It retains individual clip embeddings as temporally ordered nodes and uses learnable Gaussian weighting to regulate their affinities in the subsequent inter-modal graph. Specifically, we uniformly partition the video along the temporal axis into N densely sampled clips $\{\mathbf{V}_n\}_{n=1}^N$, and encode each clip with the shared video backbone f_v as defined in Eq. (1). This yields the clip node matrix $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_N]^\top \in \mathbb{R}^{N\times d}$. To make the graph aware of the temporal order of clips, we add a sinusoidal positional encoding $\mathbf{p}_n \in \mathbb{R}^d$ to each clip node as follows:

$$\bar{\mathbf{g}}_n = \mathbf{g}_n + \mathbf{p}_n, \quad \bar{\mathbf{G}} = [\bar{\mathbf{g}}_1, \ldots, \bar{\mathbf{g}}_N]^\top, \tag{7}$$

where $\mathbf{p}_n$ is the standard transformer sinusoidal encoding of the clip index n. Since affective cues in video are largely local in time, temporally adjacent clips are more likely to share the same emotional state. We encode this inductive bias with a Gaussian temporal kernel that down-weights edges between distant clips as follows:

$$\phi_{nn'} = \exp\left(-\frac{(n-n')^2}{2\sigma^2}\right), \quad n, n' \in [1, N]. \tag{8}$$

Here, σ is a single learnable bandwidth shared across samples and reused by the inter-modal GCN layers, providing a lightweight dataset-level temporal prior. It is optimized in log space, initialized to 3.0 in clip-index units. Combining content similarity with the temporal prior, we define the Gaussian-modulated clip-to-clip affinity as follows:

$$\tilde{\mathbf{A}}^{\text{clip}}_{nn'} = \frac{\bar{\mathbf{g}}_n^\top \bar{\mathbf{g}}_{n'}}{\|\bar{\mathbf{g}}_n\| \, \|\bar{\mathbf{g}}_{n'}\|} \cdot \phi_{nn'}, \quad n, n' \in [1, N]. \tag{9}$$

Here, the diagonal entries are set to zero to exclude self-loops. Unlike a single pooled embedding, the temporally ordered clip nodes $\bar{\mathbf{G}}$ preserve the evolution of facial behavior throughout the video, while $\phi_{nn'}$ regulates their affinities according to temporal distance. Rather than propagating a separate video-only graph, we use $\tilde{\mathbf{A}}^{\text{clip}}$ as the clip-to-clip block of the inter-modal graph, so that the clip nodes are jointly refined with the global EEG and brain-region nodes.

3.4 INTER-MODAL FUSION

In the final stage, we construct a unified inter-modal graph over the ordered video clip nodes $\{\bar{\mathbf{g}}_n\}_{n=1}^{N}$, the global EEG embedding $\mathbf{z}_e$, and the brain-region features $\{\mathbf{z}_{r_k}\}_{k=1}^{K}$. The clip and global EEG representations serve as modality nodes, while the region features serve as local nodes. The stacked node matrix is defined as follows:

$$\mathbf{H}^{\text{inter}} = [\bar{\mathbf{g}}_1, \ldots, \bar{\mathbf{g}}_N, \mathbf{z}_e, \mathbf{z}_{r_1}, \ldots, \mathbf{z}_{r_K}]^\top. \tag{10}$$

Instead of a fully connected or purely similarity-driven graph, we inject domain knowledge through a structured binary mask $\mathbf{M}^{\text{inter}}$ that retains only semantically valid connections, namely clip-to-clip edges for temporal dynamics, clip-to-EEG edges for cross-modal interaction, modality-to-region edges bridging global and local information, and inter-region edges reflecting intra-brain dependencies as follows:

$$\mathbf{M}_{ij}^{\text{inter}} = \begin{cases} 1, & \text{if } (i,j) \in \{(\bar{\mathbf{g}}_n, \bar{\mathbf{g}}_{n'}), (\bar{\mathbf{g}}_n, \mathbf{z}_e), \\ & \qquad (\bar{\mathbf{g}}_n, \mathbf{z}_{r_k}), (\mathbf{z}_e, \mathbf{z}_{r_k}), (\mathbf{z}_{r_k}, \mathbf{z}_{r_{k'}})\}, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

where $n \neq n'$, and $r_k, r_{k'} \in \mathcal{R}$ with $k \neq k'$. Following Eq. (4), affinities are computed as cosine similarities so that feature magnitude does not bias comparisons across clip, region, and global EEG nodes, where the clip-to-clip block is given by the Gaussian-modulated $\tilde{\mathbf{A}}^{\text{clip}}$ in Eq. (9). The resulting $\tilde{\mathbf{A}}^{\text{inter}}$ is restricted by $\mathbf{M}^{\text{inter}}$ and symmetrically normalized following Eq. (5) to obtain $\hat{\mathbf{A}}^{\text{inter}}$.

A two-layer GCN then integrates the temporally ordered clips, global EEG dynamics, and localized brain-region features into a unified representation, where the first layer preserves the embedding dimension and the second projects each node into the label space as follows:

$$\hat{\mathbf{H}} = \text{GCN}\left(\mathbf{H}^{\text{inter}}, \hat{\mathbf{A}}^{\text{inter}}\right) \in \mathbb{R}^{(N+1+K)\times L}, \tag{12}$$

where L denotes the number of target emotion classes. Among the updated nodes, $\hat{\mathbf{h}}_e$ is the transformed embedding of $\mathbf{z}_e$, while the video representation is obtained by averaging the updated clip nodes. The final logit is then computed as a gated combination of these two representations as:

$$\mathbf{z}_{\text{fused}} = \boldsymbol{\omega} \odot \hat{\mathbf{h}}_v + (\mathbf{1} - \boldsymbol{\omega}) \odot \hat{\mathbf{h}}_e \in \mathbb{R}^L, \;\; \hat{\mathbf{h}}_v = \frac{1}{N}\sum_{n=1}^{N} \hat{\mathbf{g}}_n \in \mathbb{R}^L \tag{13}$$

where $\boldsymbol{\omega} = \text{sigmoid}(\mathbf{w}_f)$ is a global per-class fusion gate shared across samples. This lightweight gate learns class-specific modality preferences while avoiding a sample-dependent gating network.

3.5 LOSS FUNCTIONS

To ensure consistent comparison, we adopt the following unified loss for all benchmark models as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{focal}}\mathcal{L}_{\text{focal}}, \tag{14}$$

where $\mathcal{L}_{\text{ce}}$ denotes the weighted cross-entropy loss (Cui et al., 2019) to mitigate class imbalance arising from participant's self-assessments, and $\mathcal{L}_{\text{focal}}$ denotes the focal loss to further address hard-to-classify samples.

4 EXPERIMENTAL RESULTS

4.1 IMPLEMENTATION DETAILS

4.1.1 DATASETS

We conduct experiments on two public datasets including MDMER (Yang et al., 2024) and Emognition (Saganowski et al., 2022), for which we collect paired EEG–video samples. **MDMER** dataset involves recordings from 73 participants across 32 emotion-eliciting video clips. From this dataset, we collect approximately 2.3K paired EEG–video samples, of which more than 1.8K (58 participants) are used for training and about 0.5K (15 participants) for evaluation. EEG was recorded with

18 channels, and videos were captured at 30 fps with 640×480 resolution. Ground-truth annotations were obtained from self-assessments of valence, arousal, and dominance on a 9-point scale. **Emognition** dataset contains recordings from 37 participants with affective multimedia stimuli. We construct 407 paired EEG–video samples, which are further split into 319 samples (29 participants) for training and 88 (8 participants) for evaluation. EEG was recorded with 4 channels, and videos were acquired at either 60 fps (297 samples) or 30 fps (110 samples) with 1080×1920 resolution.

Self-assessments include valence, arousal, and motivation on a 9-point scale. In our experiments, we evaluate valence and arousal recognition on both datasets. The original 9-point ratings are discretized into five categories by treating the midpoint as neutral: $\{1, 2\}$, $\{3, 4\}$, $\{5\}$, $\{6, 7\}$, and $\{8, 9\}$. We map electrodes from both datasets to canonical scalp regions defined by the international 10-20 system. Only regions containing available electrodes are instantiated. This yields five regions (*i.e.*, frontal, temporal, central, parietal, and occipital) for MDMER and two regions (*i.e.*, frontal and temporal) for Emognition. Each sample uses the complete paired EEG and video interval. Videos are divided into variable-length sequences of 32-frame clips, with frames sampled at a temporal stride of four and consecutive clips overlapping by 50%. The sampling stride is normalized to a 30 fps reference to maintain a consistent clip duration across frame rates. Facial regions are cropped and resized to 224×224 following prior studies (Zafeiriou et al., 2017; Sun et al., 2024). The EEG encoder is fine-tuned end-to-end. Additional dataset details are provided in Appendix A.7.2, while the clip sampling configuration is analyzed in Appendix A.5.

4.1.2 Environments

All experiments use the same subject-independent splits, training budget, loss formulation, and evaluation metrics to ensure a fair comparison. For the benchmark evaluation, all networks are trained for 100 epochs with a batch size of 12, while model-specific learning rates and weight decay follow their recommended configurations. Existing multi-modal baselines retain their original architectures. We report Accuracy, Unweighted Average Recall (UAR), and Weighted F1-score (W-F1). We set $\lambda_{\text{ce}} = 0.9$ and $\lambda_{\text{focal}} = 0.1$ for all networks. More details are provided in the Appendix A.7.1.

4.2 Benchmark Results

Table 1 presents the results on Emognition and MDMER datasets under unimodal and multi-modal settings. Each baseline follows its training configuration to ensure a fair comparison. The relative strength of the two modalities varies across datasets. On Emognition, the average accuracy of the EEG backbones ranges from 27.6% to 35.3%, exceeding that of the video backbones, which ranges from 23.9% to 30.7%. In contrast, video provides stronger representations on MDMER, achieving 29.5-33.4% average accuracy compared with 25.1-30.2% for EEG. This dataset-dependent behavior indicates that neither modality is consistently dominant and motivates structured fusion that exploits their complementary information. Existing multi-modal methods do not consistently improve upon the strongest unimodal models. EAV (Lee et al., 2024) and Milmer (Wang et al., 2026) obtain average accuracies of 27.8% and 29.5% on the Emognition and 33.3% and 31.7% on the MDMER, respectively. These results remain below CBraMod on the Emognition at 35.3% and VideoMAE on MDMER at 33.4%, showing that combining modalities does not guarantee improved performance.

In contrast, EVER consistently outperforms the evaluated unimodal and multi-modal baselines across both datasets. The AdaMAE+CBraMod configuration achieves the highest average accuracy of 38.6% on the Emognition, exceeding the strongest baseline by 3.3 percentage points. On the MDMER, VideoMAE+CBraMod reaches 35.5%, providing a 2.1% improvement over the best baseline. VideoMAE+CBraMod also achieves the best balanced metrics, obtaining 35.3% UAR and 35.9% W-F1 on the Emognition. These scores improve upon the strongest baselines by 6.6 and 5.3 percentage points, respectively. This indicates that the proposed fusion improves class-balanced recognition rather than relying primarily on dominant classes. The two backbone configurations exhibit complementary behavior. VideoMAE+CBraMod performs best on UAR and W-F1, whereas AdaMAE+CBraMod achieves the highest accuracy. Nevertheless, both configurations consistently outperform the competing unimodal and multi-modal approaches. The gains arise from preserving anatomically structured EEG representations and temporally ordered video features, followed by node-level reasoning across the two modalities. Additional analyses of statistical reliability and cross-modal complementarity, together with comparisons against additional EEG architectures, are provided in Appendices A.3, A.4, and A.6.

Table 1: **Performance comparison on the Emognition and MDMER datasets.** We report Valence, Arousal, and their average score in terms of Accuracy (Acc), UAR, and W-F1.

Modality	Model	Target	Emognition			MDMER		
			Acc	UAR	W-F1	Acc	UAR	W-F1
Video	ViViT (Arnab et al., 2021)	Valence	0.182	0.166	0.174	0.300	0.218	0.278
		Arousal	0.296	0.228	0.248	0.290	0.230	0.273
		Average	0.239	0.197	0.211	0.295	0.224	0.274
	Swin (Liu et al., 2021)	Valence	0.295	0.235	0.257	0.349	0.246	0.308
		Arousal	0.239	0.196	0.204	0.292	0.201	0.237
		Average	0.267	0.215	0.230	0.320	0.223	0.273
	TSF (Bertasius et al., 2021)	Valence	0.295	0.222	0.209	0.309	0.202	0.212
		Arousal	0.261	0.222	0.218	0.307	0.216	0.237
		Average	0.278	0.222	0.214	0.308	0.209	0.224
	VideoMAE (Tong et al., 2022)	Valence	0.284	0.260	0.274	0.355	0.229	0.266
		Arousal	0.307	0.256	0.285	0.313	0.211	0.178
		Average	0.296	0.258	0.280	0.334	0.220	0.222
	AdaMAE (Bandara et al., 2023)	Valence	0.307	0.230	0.214	0.359	0.233	0.285
		Arousal	0.307	0.266	0.244	0.305	0.218	0.235
		Average	0.307	0.248	0.229	0.332	0.225	0.260
EEG	ST-Transformer (Song et al., 2021)	Valence	0.235	0.214	0.187	0.236	0.174	0.151
		Arousal	0.400	0.320	0.332	0.265	0.196	0.251
		Average	0.318	0.267	0.259	0.251	0.185	0.201
	BIOT (Yang et al., 2023)	Valence	0.341	0.334	0.329	0.248	0.203	0.243
		Arousal	0.306	0.236	0.283	0.255	0.221	0.258
		Average	0.324	0.285	0.306	0.252	0.212	0.251
	LaBraM (Jiang et al., 2024)	Valence	0.259	0.200	0.106	0.309	0.189	0.172
		Arousal	0.294	0.273	0.171	0.288	0.200	0.148
		Average	0.276	0.237	0.139	0.299	0.195	0.160
	CBraMod (Wang et al., 2025)	Valence	0.365	0.342	0.366	0.320	0.211	0.232
		Arousal	0.341	0.232	0.218	0.284	0.206	0.206
		Average	0.353	0.287	0.292	0.302	0.208	0.219
	REVE (El Ouahidi et al., 2025)	Valence	0.294	0.243	0.263	0.284	0.189	0.213
		Arousal	0.259	0.233	0.218	0.282	0.220	0.256
		Average	0.276	0.238	0.240	0.283	0.205	0.234
Video+EEG	EAV (Lee et al., 2024)	Valence	0.284	0.275	0.257	0.311	0.190	0.188
		Arousal	0.273	0.246	0.168	0.355	0.263	0.272
		Average	0.278	0.260	0.212	0.333	0.226	0.230
	Milmer (Wang et al., 2026)	Valence	0.227	0.233	0.165	0.338	0.218	0.222
		Arousal	0.364	0.273	0.268	0.296	0.207	0.224
		Average	0.295	0.253	0.216	0.317	0.212	0.223
	Ours (VideoMAE+CBraMod)	Valence	0.398	0.376	0.372	0.380	0.251	0.316
		Arousal	0.352	0.329	0.345	0.330	0.225	0.247
		Average	0.375	**0.353**	**0.359**	**0.355**	**0.238**	**0.281**
	Ours (AdaMAE+CBraMod)	Valence	0.364	0.313	0.297	0.349	0.230	0.294
		Arousal	0.409	0.321	0.331	0.330	0.230	0.256
		Average	**0.386**	0.317	0.314	0.339	0.230	0.275

Bold: The best, Underline: The second-best

4.3 Ablation Study

We report the mean of valence and arousal scores. All fusion ablations use identical AdaMAE and CBraMod backbones to isolate the fusion mechanism. Additional analyses of robustness to test-time EEG corruption and model size are provided in Appendices A.1 and A.2.

4.3.1 Component Analysis

Table 2 evaluates the components of EVER on the Emognition dataset. The unimodal video and EEG models achieve accuracies of 30.7% and 35.3%, respectively. Introducing BAR-GCN region nodes into the inter-modal graph with a pooled video representation increases the accuracy to 37.5%, already outperforming both unimodal baselines. Clip nodes retain clip embeddings in the inter-modal graph, whereas the VCG prior further equips them with positional encoding and Gaussian temporal modulation. Simply introducing video clips as additional graph nodes without temporal structure reduces the accuracy to 35.2%, suggesting that unordered clip representations may introduce temporal ambiguity. Incorporating the VCG prior, which combines positional encoding with learnable Gaussian temporal modulation, increases the accuracy to 38.6% and the UAR to 31.7%.

4.3.2 Effect of Anatomical Structure

Table 3 evaluates BAR-GCN on MDMER dataset, whose montage spans five anatomical regions. Replacing the anatomical partition with a random channel-to-region partition of the same size retains

Table 2: **Ablation study of EVER components on the Emognition dataset.**

Modality	BAR-GCN	Clip nodes	VCG	Emognition		
				Acc	UAR	W-F1
Video				0.307	0.248	0.229
EEG				0.353	0.287	0.292
Video + EEG	✓			0.375	0.294	**0.315**
	✓	✓		0.352	0.288	0.307
	✓	✓	✓	**0.386**	**0.317**	0.314

Bold: The best, Underline: The second-best

Table 3: **Ablation study of the BAR-GCN on the MDMER dataset.**

Method	Variant	MDMER		
		Acc	UAR	W-F1
Ours	Full	**0.339**	**0.230**	**0.275**
	Random region partition	0.326	0.208	0.185
	w/o inter-region edges	0.318	0.205	0.196
	w/o $\mathbf{M}^{\text{local}}$	0.311	0.198	0.154
	Pooled region features	0.298	0.206	0.206

Bold: The best, Underline: The second-best

Table 4: **Ablation of the VCG temporal prior on the Emognition dataset.** $\mathbf{p}_n$ denotes the positional encoding and $\phi_{nn'}$ denotes the learnable Gaussian prior.

Method	Variant	Emognition		
		Acc	UAR	W-F1
Ours (VCG prior)	Full	**0.386**	**0.317**	**0.314**
	w/o $\mathbf{p}_n$	0.307	0.261	0.245
	w/o $\phi_{nn'}$	0.290	0.249	0.187

Bold: The best, Underline: The second-best

Table 5: **Ablation study of the fusion strategy on the Emognition dataset.** "Pooled modality nodes" averages the clip and region features before inter-modal fusion.

Fusion	Variant	Emognition		
		Acc	UAR	W-F1
w/o graph	Concatenation	0.290	0.233	0.218
	Cross-attention	0.308	0.272	0.286
	FiLM	0.308	0.308	0.286
GCN	Pooled modality nodes	0.285	0.200	0.127
	Ours (Inter-modal)	**0.386**	**0.317**	**0.314**

Bold: The best, Underline: The second-best

an accuracy of 32.6% but reduces W-F1 from 27.5% to 18.5%. This result suggests that grouping channels alone can preserve predictive performance, whereas anatomically meaningful assignments are important for balanced class recognition. Removing the intra-region mask $\mathbf{M}^{\text{local}}$ and the inter-region edges further reduces the accuracy to 31.1% and 31.8%, respectively. Pooling the region features before constructing the inter-modal graph produces the lowest accuracy of 29.8%. Together, these results support the contributions of anatomical grouping, constrained intra-region interaction, inter-region propagation, and fine-grained node preservation to cross-subject emotion recognition.

4.3.3 Analysis of the VCG Prior

Table 4 evaluates the positional encoding $\mathbf{p}_n$ and Gaussian temporal weight $\phi_{nn'}$ within the VCG prior. Removing either component reduces the accuracy from 38.6% to 30.7% and 29.0%, respectively, with decreases in UAR and W-F1. These results suggest that the full VCG benefits from jointly combining temporal-order information from $\mathbf{p}_n$ with distance-aware modulation from $\phi_{nn'}$, rather than from either component alone.

4.3.4 Fusion Strategy

Table 5 provides a controlled comparison using identical EEG-video backbones, fine-tuning settings, and training objectives across all fusion strategies. Concatenation achieves 29.0% accuracy, while Cross-attention and FiLM methods both reach 30.8%. Applying a GCN only to globally pooled modality nodes yields 28.5%, indicating that graph convolution alone does not account for the performance improvement. In contrast, operating the graph over fine-grained video clip and EEG region nodes increases the accuracy to 38.6%. This comparison shows that the main benefit arises from structured node-level interaction rather than the fusion operator alone.

5 Conclusion

In this paper, we proposed EVER for EEG–video emotion recognition. EVER combines anatomy-aware EEG modeling with temporally structured video reasoning through an inter-modal graph. We also conducted an extensive benchmark across two public datasets and twelve representative models. Experimental results show that EVER outperforms unimodal and conventional multi-modal approaches. The variable-length clip design improves temporal coverage but increases computation in proportion to the video duration. Future work will explore efficient clip selection to enable evaluation over additional subject-independent splits.

AI USE STATEMENT

Generative AI tools were used after the initial manuscript draft was completed to improve the clarity and presentation of the writing and to assist with the retrieval and discovery of relevant prior work. All AI-assisted edits and retrieved references were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the methodological details required to reproduce EVER in Sec. 3, with additional training, optimization, and implementation settings in Appendix A.7. Additional analyses of robustness, model size, statistical reliability, modality complementarity, video clip configuration, and EEG backbones are provided in Appendices A.1–A.6. All experiments use publicly available datasets and pretrained models, and the source code will be released upon acceptance.

REFERENCES

Joseph Aina, Oluwatunmise Akinniyi, Md Mahmudur Rahman, Valerie Odero-Marah, and Fahmi Khalifa. A hybrid learning-architecture for mental disorder detection using emotion recognition. *IEEE Access*, 12:91410–91425, 2024.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pp. 6836–6846, 2021.

Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 14507–14517, 2023.

Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(9):5826–5846, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2021.

M Kalpana Chowdary, Tu N Nguyen, and D Jude Hemanth. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, 35(32):23311–23328, 2023.

Karina Cortiñas-Lorenzo and Gerard Lacey. Toward explainable affective computing: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13101–13121, 2023.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pp. 9268–9277, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Guanglong Du, Jinshao Su, Linlin Zhang, Kang Su, Xueqian Wang, Shaohua Teng, and Peter Xiaoping Liu. A multi-dimensional graph convolution network for eeg emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg-adapting to any setup with large-scale pretraining on 25,000 subjects. *Advances in Neural Information Processing Systems*, 38:22541–22577, 2025.

Reham A Elsheikh, MA Mohamed, Ahmed Mohamed Abou-Taleb, and Mohamed Maher Ata. Improved facial emotion recognition model based on a novel deep convolutional structure. *Scientific Reports*, 14(1):29050, 2024.

Yue Gao, Xiangling Fu, Tianxiong Ouyang, and Yi Wang. Eeg-gcn: Spatio-temporal and self-adaptive graph convolutional networks for single and multi-view eeg-based emotion recognition. *IEEE Signal Processing Letters*, 29:1574–1578, 2022.

Uwe Herwig, Peyman Satrapi, and Carlos Schönfeldt-Lecuona. Using the international 10-20 eeg system for positioning of transcranial magnetic stimulation. *Brain topography*, 16(2):95–99, 2003.

Miho Iwasaki and Yasuki Noguchi. Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements. *Scientific Reports*, 6(1):22049, 2016.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Ming Jin, Changde Du, Huiguang He, Ting Cai, and Jinpeng Li. Pgcn: Pyramidal graph convolutional network for eeg emotion recognition. *IEEE Transactions on Multimedia*, 26:9070–9082, 2024.

Stamos Katsigiannis and Naeem Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE journal of Biomedical and Health Informatics*, 22(1):98–107, 2017.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2017.

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, 2011.

Min-Ho Lee, Adai Shomanov, Balgyn Begim, Zhuldyz Kabidenova, Aruna Nyssanbay, Adnan Yazici, and Seong-Whan Lee. Eav: Eeg-audio-video dataset for emotion recognition in conversational contexts. *Scientific Data*, 11(1):1026, 2024.

Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. Vbh-gnn: Variational bayesian heterogeneous graph neural networks for cross-subject emotion recognition. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

Greta Macorig, Arielle Crespel, Annacarmen Nilo, Ngoc Phuong Loc Tang, Mariarosaria Valente, Gian Luigi Gigli, and Philippe Gélisse. Benign eeg variants in the sleep–wake cycle: A prospective observational study using the 10–20 system and additional electrodes. *Neurophysiologie Clinique*, 51(3):233–242, 2021.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Deng Pan, Haohao Zheng, Feifan Xu, Yu Ouyang, Zhe Jia, Chu Wang, and Hong Zeng. Msfr-gcn: A multi-scale feature reconstruction graph convolutional network for eeg emotion and cognition recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3245–3254, 2023.

Yong Peng, Honggang Liu, Wanzeng Kong, Feiping Nie, Bao-Liang Lu, and Andrzej Cichocki. Joint eeg feature transfer and semisupervised cross-subject emotion recognition. *IEEE Transactions on Industrial Informatics*, 19(7):8104–8115, 2022.

AJ Piergiovanni and Michael Ryoo. Temporal gaussian mixture layer for videos. In *International Conference on Machine learning*, pp. 5152–5161. PMLR, 2019.

Stanisław Saganowski, Joanna Komoszyńska, Maciej Behnke, Bartosz Perz, Dominika Kunc, Bartłomiej Klich, Łukasz D Kaczmarek, and Przemysław Kazienko. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific Data*, 9(1):158, 2022.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

Gourav Siddhad, Anmol Gupta, Debi Prosad Dogra, and Partha Pratim Roy. Efficacy of transformer networks for classification of raw eeg data. *Biomedical Signal Processing and Control*, 87:105488, 2024.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion*, 108:102382, 2024.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 35:10078–10093, 2022.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 399–417, 2018.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 37:36314–36341, 2024.

Zaitian Wang, Jian He, Yu Liang, Xiyuan Hu, Tianhao Peng, Kaixin Wang, Jiakai Wang, Chenlong Zhang, Weili Zhang, Shuang Niu, et al. Milmer: a framework for multiple instance learning based multimodal emotion recognition. *Neurocomputing*, pp. 134368, 2026.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3289–3299, 2020.

Pei Yang, Niqi Liu, Xinge Liu, Yezhi Shu, Wenqi Ji, Ziqi Ren, Jenny Sheng, Minjing Yu, Ran Yi, Dan Zhang, et al. A multimodal dataset for mixed emotion recognition. *Scientific Data*, 11(1): 847, 2024.

Mengqing Ye, CL Philip Chen, and Tong Zhang. Hierarchical dynamic graph convolutional network with interpretability for eeg-based emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proc. of Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 34–41, 2017.

Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1034–1047, 2021.

Shiqing Zhang, Xianzhang Pan, Yueli Cui, Xiaoming Zhao, and Limei Liu. Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access*, 7:32297–32304, 2019.

Xiaowei Zhang, Jinyong Liu, Jian Shen, Shaojie Li, Kechen Hou, Bin Hu, Jin Gao, and Tong Zhang. Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine. *IEEE Transactions on Cybernetics*, 51(9):4386–4399, 2020.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *Proc. of Int'l Conf. on Learning Representations (ICLR)*, 2023.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the International Conference on Machine Learning*, pp. 27268–27286, 2022.

A Appendix

In this Appendix, we provide additional analyses, extended comparisons, and implementation details that could not be included in the main manuscript. Specifically, we first evaluate the robustness of EVER under test-time EEG corruption and report the computational cost of all benchmarked models. We further examine the statistical reliability of the benchmark through bootstrap confidence intervals, and analyze the complementarity between the video and EEG branches at the sample level. Additional results include a comparison with six transformer architectures for multivariate time series adapted to EEG classification, together with the detailed experimental setup and the annotation distributions of the Emognition (Saganowski et al., 2022) and MDMER (Yang et al., 2024) datasets. The source code will be made publicly available upon acceptance.

A.1 Robustness to EEG Corruption

To evaluate robustness under unreliable EEG signals, we conduct a test-time corruption experiment on the Emognition dataset (Saganowski et al., 2022) without retraining the models. For each EEG input $\mathbf{x}$, additive Gaussian noise is applied as follows:

$$\mathbf{x}' = \mathbf{x} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, (\lambda \sigma_{\mathbf{x}})^2\right), \tag{15}$$

where $\sigma_{\mathbf{x}}$ denotes the standard deviation of the input EEG tensor and $\lambda \in \{0, 0.1, 0.2, 0.3, 0.5\}$ controls the corruption level. The same corrupted EEG signal is provided to both the global CBraMod (Wang et al., 2025) pathway and the STFT-based regional pathway, while the video input remains unchanged. We compare the EEG-only CBraMod model with the AdaMAE- (Bandara et al., 2023) and VideoMAE-based (Tong et al., 2022) versions of EVER using Accuracy, UAR, and W-F1 averaged over valence and arousal.

As shown in Table A1, CBraMod decreases from 0.353 to 0.294 in Accuracy as λ increases from 0 to 0.5, corresponding to a degradation of 0.059 or 16.7% of its clean performance. In contrast,

Table A1: **Robustness to test-time EEG corruption on the Emognition dataset.** Each entry reports Acc/UAR/W-F1 averaged over valence and arousal.

Method	Backbone	Noise level λ					Δ Acc
		0.0	0.1	0.2	0.3	0.5	
EEG-only	CBraMod	0.353/0.287/0.292	0.329/0.261/0.271	0.318/0.253/0.261	0.306/0.245/0.253	0.294/0.241/0.244	−0.059
Video + EEG (Ours)	AdaMAE	0.386/0.317/0.314	0.365/0.297/0.292	0.365/0.297/0.292	0.347/0.281/0.264	0.341/0.272/0.254	−0.045
	VideoMAE	0.375/**0.353/0.359**	0.366/**0.350/0.348**	0.359/**0.347/0.335**	0.359/**0.347/0.334**	0.335/**0.315/0.304**	**−0.040**

A smaller absolute Δ Acc indicates better robustness.

Table A2: **Model size and average performance on the Emognition and MDMER datasets.** All performance values are averaged over valence and arousal.

Modality	Model	Params	Emognition			MDMER		
			Acc	UAR	W-F1	Acc	UAR	W-F1
Video	ViViT (Arnab et al., 2021)	88.7M	0.239	0.197	0.211	0.295	0.224	0.274
	Swin (Liu et al., 2021)	87.6M	0.267	0.215	0.230	0.320	0.223	0.273
	TSF (Bertasius et al., 2021)	121.3M	0.278	0.222	0.214	0.308	0.209	0.224
	VideoMAE (Tong et al., 2022)	86.4M	0.296	0.258	0.280	0.334	0.220	0.222
	AdaMAE (Bandara et al., 2023)	87.6M	0.307	0.248	0.229	0.332	0.225	0.260
EEG	ST-Transformer (Song et al., 2021)	2.6M	0.318	0.267	0.259	0.251	0.185	0.201
	BIOT (Yang et al., 2023)	3.2M	0.324	0.285	0.306	0.252	0.212	0.251
	LaBraM (Jiang et al., 2024)	5.8M	0.276	0.237	0.139	0.299	0.195	0.160
	CBraMod (Wang et al., 2025)	21.3M	0.353	0.287	0.292	0.302	0.208	0.219
	REVE (El Ouahidi et al., 2025)	111.9M	0.276	0.238	0.240	0.283	0.205	0.234
Video+EEG	EAV (Lee et al., 2024)	85.8M	0.278	0.260	0.212	0.333	0.226	0.230
	Milmer (Wang et al., 2026)	1457.6M	0.295	0.253	0.216	0.317	0.212	0.223
	Ours(VideoMAE+CBraMod)	135.8M	<u>0.375</u>	**0.353**	**0.359**	**0.355**	**0.238**	**0.281**
	Ours(AdaMAE+CBraMod)	136.9M	**0.386**	<u>0.317</u>	<u>0.314</u>	<u>0.339</u>	<u>0.230</u>	<u>0.275</u>

Bold: The best, <u>Underline</u>: The second-best

both versions of EVER degrade less, by 0.045 for AdaMAE and 0.040 for VideoMAE, and the VideoMAE-based fusion also maintains the highest UAR and W-F1 across all corruption levels. The margin over the EEG-only model therefore widens with the corruption level, from 0.033 to 0.047 in Accuracy for AdaMAE and from 0.022 to 0.041 for VideoMAE. Under the strongest corruption, both variants remain above the best video-only baseline (*i.e.*, AdaMAE) of 0.307, showing that the fused prediction collapses to neither the degraded EEG branch nor the video branch alone. Since the fusion gate is shared across samples, this robustness is attributed to multi-modal redundancy and to the inter-modal graph, in which the intact clip nodes continue to propagate information to the corrupted EEG nodes, rather than to sample-adaptive modality reweighting.

A.2 Model Size and Accuracy

As shown in Table A2, EVER provides a favorable trade-off between model capacity and recognition performance. The two EVER variants contain 135.8M and 136.9M parameters, only about 9.3% of Milmer's 1457.6M, and both outperform Milmer and EAV across all metrics on both datasets. Since the paired backbones already account for 107.7M and 108.9M, the proposed modules add approximately 28M parameters, suggesting that parameter count alone does not explain the improvement. Most EEG models are compact, ranging from 2.6M to 21.3M, with REVE as the main exception at 111.9M, which makes CBraMod a parameter-efficient global EEG backbone for fusion.

The relative effectiveness of the two modalities varies across datasets. On Emognition, the EEG-only CBraMod model reaches 0.353 in average Accuracy and exceeds the best video-only baseline at 0.307, whereas on MDMER the best video model achieves 0.334 against 0.302 for the best EEG model. This difference suggests that the informativeness of each modality depends on dataset-specific factors such as recording quality, subject variability, stimulus design, and the strength of observable facial responses. Relying on a single modality is therefore unlikely to provide consistently strong performance across acquisition conditions. EVER addresses this limitation by jointly exploiting complementary physiological and visual cues, and Ours (VideoMAE+CBraMod) improves over its own video and EEG backbones by 0.079 and 0.022 on Emognition and by 0.021 and 0.053 on MDMER. Thus, the gain is obtained against whichever modality is weaker on a given dataset, showing that the fusion compensates for dataset-dependent weaknesses rather than inheriting the behavior of a single branch.

Table A3: **Bootstrap 95% confidence intervals on the Emognition and MDMER datasets.** Intervals are percentile bootstrap CIs obtained from $B = 10{,}000$ resamples, averaged over valence and arousal.

Modality	Model	Emognition			MDMER		
		Acc	UAR	W-F1	Acc	UAR	W-F1
Video	ViViT	0.239 [.186,.298]	0.197 [.150,.247]	0.211 [.156,.268]	0.295 [.266,.324]	0.224 [.205,.243]	0.274 [.244,.304]
	Swin	0.267 [.208,.326]	0.215 [.171,.261]	0.230 [.182,.287]	0.320 [.291,.349]	0.223 [.204,.243]	0.273 [.244,.303]
	TSF	0.278 [.213,.349]	0.222 [.183,.259]	0.214 [.157,.280]	0.308 [.277,.339]	0.209 [.189,.230]	0.224 [.193,.255]
	VideoMAE	0.296 [.225,.367]	0.258 [.205,.312]	0.280 [.215,.350]	0.334 [.303,.365]	0.220 [.215,.225]	0.222 [.196,.250]
	AdaMAE	0.307 [.236,.378]	0.248 [.203,.292]	0.229 [.172,.296]	0.332 [.303,.360]	0.225 [.209,.241]	0.260 [.233,.287]
EEG	ST-Transformer	0.318 [.256,.381]	0.267 [.215,.322]	0.259 [.196,.326]	0.251 [.225,.278]	0.185 [.163,.209]	0.201 [.177,.227]
	BIOT	0.324 [.250,.404]	0.285 [.219,.353]	0.306 [.229,.385]	0.252 [.223,.282]	0.212 [.181,.247]	0.251 [.221,.282]
	LaBraM	0.276 [.196,.356]	0.237 [.210,.263]	0.139 [.082,.208]	0.299 [.269,.329]	0.195 [.185,.205]	0.160 [.135,.187]
	CBraMod	0.353 [.279,.427]	0.287 [.229,.348]	0.292 [.216,.369]	0.302 [.276,.328]	0.208 [.193,.223]	0.219 [.193,.246]
	REVE	0.276 [.208,.350]	0.238 [.182,.296]	0.240 [.172,.314]	0.283 [.254,.312]	0.205 [.182,.232]	0.234 [.205,.263]
Video+EEG	EAV	0.278 [.202,.354]	0.260 [.199,.324]	0.212 [.144,.287]	0.333 [.301,.365]	0.226 [.211,.241]	0.230 [.198,.262]
	Milmer	0.295 [.219,.371]	0.253 [.194,.314]	0.216 [.147,.292]	0.317 [.286,.348]	0.212 [.196,.228]	0.223 [.193,.254]
	Ours(VideoMAE+CBraMod)	0.375 [.295,.455]	**0.353** [.272,.437]	**0.359** [.279,.440]	**0.355** [.322,.387]	**0.238** [.220,.255]	**0.281** [.248,.313]
	Ours(AdaMAE+CBraMod)	**0.386** [.312,.460]	0.317 [.266,.368]	0.314 [.238,.395]	0.339 [.309,.369]	0.230 [.211,.248]	0.275 [.246,.305]

Bold: The best, Underline: The second-best

A.3 Statistical Reliability of the Benchmark

Because paired EEG-video datasets are small and our splits are subject-independent, aggregate scores carry non-negligible sampling variance. To quantify test-set sampling uncertainty for each fixed trained model, we compute percentile bootstrap confidence intervals from $B = 10{,}000$ resamples of the test predictions in Table A3, where all models are evaluated on identical resamples.

The interval widths follow the test sizes of the two datasets. On Emognition with 88 evaluation samples, the Accuracy interval of Ours (VideoMAE+CBraMod) spans 0.160, whereas the corresponding interval on MDMER with about 0.5K samples spans 0.065. The Emognition intervals of the strongest models therefore overlap, and we do not claim that individual differences on this dataset are statistically separated. On MDMER, the Accuracy interval of Ours (VideoMAE+CBraMod) is disjoint from those of ST-Transformer, BIOT, and REVE, and its W-F1 interval is disjoint from those of ST-Transformer, LaBraM, and CBraMod. The intervals of the strongest video baselines remain partially overlapping, so the ranking among the top models should be interpreted as consistent rather than significant. We regard this consistency as the more informative evidence, since EVER attains the best or second-best estimate in all six metric and dataset combinations across two independent video backbones, whereas no baseline achieves this consistency.

A.4 Complementarity of EEG and Video

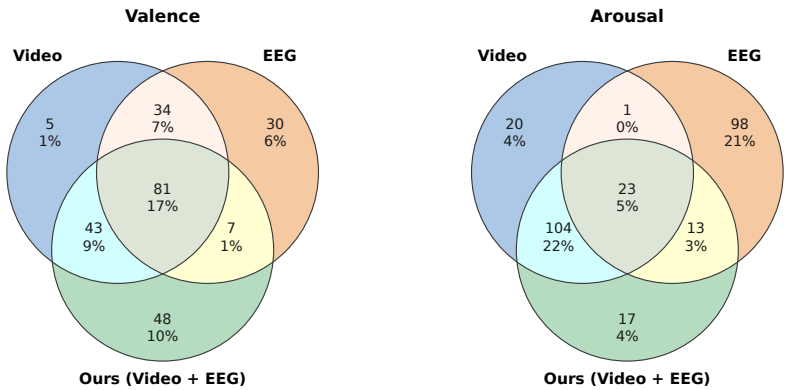


Figure A1: **Complementary correctness of the video, EEG, and fused predictions on the MDMER dataset.** Each region reports the number of test samples and its share of the evaluation set. Circle sizes are fixed for legibility and are not area-proportional.

To examine whether the two modalities succeed on the same samples, we compare the correctly classified sets of the video-only model, the EEG-only CBraMod model, and Ours (Video + EEG) on the MDMER test split in Figure A1. The complementary rate, defined as the fraction of samples solved by exactly one modality, reaches 18% for valence and 49% for arousal. On arousal, video

Table A4: **Effect of the clip sampling configuration on the Emognition dataset.** Both settings use the AdaMAE backbone with a learning rate of 1×10^{-4} and a weight decay of 0.05.

Backbone	Frame interval	Overlap	Emognition		
			Acc	UAR	W-F1
AdaMAE	2	–	0.278	0.235	0.212
	4 (ours)	0.5	**0.307**	**0.248**	**0.229**

Table A5: **Comparison with transformer architectures for multivariate time series, adapted to EEG classification.** We use the classification-adapted implementations from the Medformer benchmark (Wang et al., 2024). All values are averaged over valence and arousal, and all models are trained under the same subject-independent splits and training budget as the main benchmark.

Modality	Model	Emognition			MDMER		
		Acc	UAR	W-F1	Acc	UAR	W-F1
EEG	PatchTST (Nie et al., 2023)	0.259	0.210	0.219	0.281	0.204	<u>0.266</u>
	Informer (Zhou et al., 2021)	<u>0.318</u>	0.241	0.208	0.300	0.210	0.215
	FEDformer (Zhou et al., 2022)	0.282	<u>0.270</u>	<u>0.275</u>	0.264	0.195	0.244
	Crossformer (Zhang & Yan, 2023)	0.282	0.214	0.221	<u>0.322</u>	0.205	0.184
	Medformer (Wang et al., 2024)	0.312	0.223	0.206	0.318	0.212	0.230
	EEG-Transformer (Siddhad et al., 2024)	0.300	0.212	0.171	0.302	<u>0.217</u>	0.200
Video+EEG	**Ours (VideoMAE+CBraMod)**	0.375	**0.353**	**0.359**	**0.355**	**0.238**	**0.281**
	Ours (AdaMAE+CBraMod)	**0.386**	0.317	0.314	0.339	0.230	0.275

Bold: The best, <u>Underline</u>: The best among the additional EEG models

Table A6: **Model-specific learning rate and weight decay.** All other training settings are shared across models.

Modality	Model	LR	Weight decay
Video	ViViT, Swin, TSF	2×10^{-5}	0.02
	VideoMAE, AdaMAE	1×10^{-4}	0.05
EEG	ST-Transformer, BIOT, LaBraM, CBraMod, REVE	1×10^{-4}	0.05
	PatchTST, Informer, FEDformer, Crossformer, Medformer, EEG-Transformer	1×10^{-4}	0.05
Video+EEG	**Ours (VideoMAE/AdaMAE + CBraMod)**	1×10^{-4}	0.05

and EEG each solve about a quarter of the test set exclusively while agreeing on only 5%, so any unimodal model forfeits a large portion of the available evidence.

Ours (Video + EEG) exploits this evidence in two ways. First, 48 samples for valence and 17 for arousal are classified correctly although both unimodal models fail, which cannot be explained by selecting the better modality prediction. These cases are consistent with complementary information being exploited through joint refinement of region-level EEG and temporally ordered video nodes. Second, the agreement between the branches is rarely overturned, as 23 of the 24 samples solved by both modalities remain correct on arousal. The net effect improves over the best single modality from 31.3% to 33.0% on arousal.

The recovery of exclusive evidence follows the relative strength of the modalities on this dataset, where video is the stronger branch and 104 of the 124 video-exclusive samples are retained. Importantly, this preference does not come at the cost of the weaker branch. The fused model still solves 17 arousal and 48 valence samples that neither branch solves alone, so the shared gate reweights the modalities at the dataset level while the inter-modal graph continues to exploit EEG-specific evidence. Accessing this evidence is a matter of representation rather than capacity, since Milmer fuses globally aggregated features with roughly ten times the parameters of EVER (Table A2) yet reaches 0.317 on MDMER, below the 0.334 of the video-only VideoMAE backbone.

A.5 Video Clip Configuration

All video models in the benchmark, including EVER, sample clips with a frame interval of four and an overlap ratio of 0.5. Table A4 compares this setting against non-overlapping clips with a frame interval of two on the AdaMAE backbone under an identical optimization setting. The two configurations advance by the same hop of 64 frames and therefore yield the same number of clips

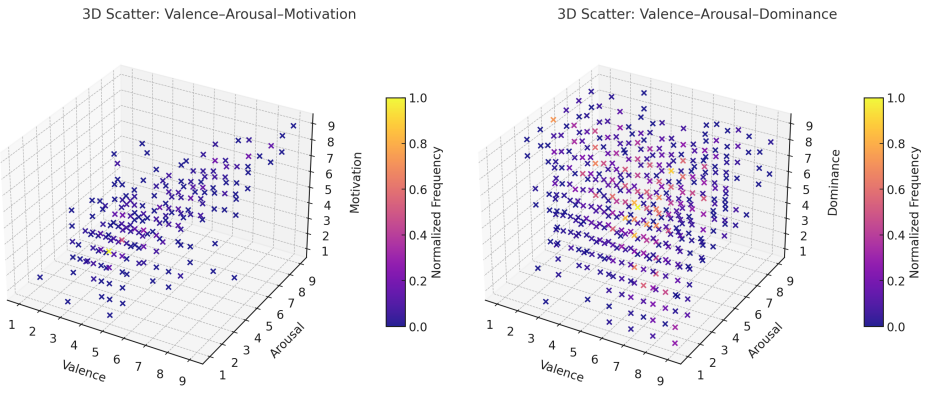
and the same overall coverage of a recording, but each of our clips spans twice the temporal extent and shares half of its content with its neighbors. This widens the temporal context available to a single clip and keeps adjacent clips continuous rather than disjoint, improving Accuracy from 0.278 to 0.307 with consistent gains in UAR and W-F1 at no additional computational cost. Thus, we adopt it for every video backbone, so that the reported gains of EVER are obtained against baselines evaluated under their stronger setting.

A.6 Additional EEG Backbones

The EEG models in the main benchmark are either EEG foundation models or transformers designed for general EEG decoding. To examine whether the advantage of EVER persists against a different family of EEG encoders, we additionally benchmark six transformer architectures for multivariate time series, namely PatchTST (Nie et al., 2023), Informer (Zhou et al., 2021), FEDformer (Zhou et al., 2022), Crossformer (Zhang & Yan, 2023), Medformer (Wang et al., 2024), and EEG-Transformer (Siddhad et al., 2024). The first four were originally proposed for long-term forecasting, whereas the latter two target medical and EEG signals directly. We adopt the classification-adapted implementations released with the Medformer benchmark (Wang et al., 2024), which reformulates these forecasting architectures for medical time-series classification on EEG and ECG recordings. Since they encode multi-channel EEG through the same input interface as our benchmark, they can be trained under identical subject-independent splits, loss formulation, and training budget, and therefore serve as an independent set of EEG-only references obtained without any modification of our evaluation protocol.

As shown in Table A5, these models perform within the range of the EEG backbones reported in the main paper. On Emognition, their average Accuracy ranges from 0.259 to 0.318, and on MDMER from 0.264 to 0.322. Crossformer reaches 0.322 on MDMER and thereby exceeds the strongest EEG model of the main benchmark at 0.302, which indicates that these architectures are competitive EEG encoders rather than weak references. Nevertheless, all six remain below both versions of EVER on every metric and dataset. The gap is largest for the class-balanced metrics on Emognition, where EVER improves UAR from 0.270 to 0.353 and W-F1 from 0.275 to 0.359 relative to the best of these models. Since the six models span frequency-domain, patch-based, and cross-dimensional attention designs, their consistent shortfall suggests that the limitation of the EEG-only setting is the absence of the complementary visual modality rather than the choice of EEG architecture. Together with the corruption experiment in Sec. A.1, this supports that the benefit of Ours (Video + EEG) is robust to both the quality of the EEG signal and the design of the EEG encoder. All six models use the classification implementations of the Medformer codebase (Wang et al., 2024) and are trained with the same learning rate and weight decay as the EEG backbones of the main benchmark (Table A6), under an identical optimizer, schedule, loss, and training budget.

A.7 Additional Implementation Details



Emognition dataset MDMER dataset

Figure A2: Distribution of self-assessment annotations (*i.e.*, ground-truth scores) on the Emognition and MDMER datasets.

A.7.1 EXPERIMENTAL SETUP

Although we have already described the environment setup in the main manuscript, we provide additional details of the benchmark configuration. We employ the AdamW optimizer with $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. The learning rate follows a linear warmup over the first 10% of the total training steps, after which it is decayed by a factor of 0.75 every ten epochs starting from the tenth epoch, implemented through a LambdaLR scheduler. All networks are trained for 100 epochs with a batch size of 12, and training is terminated early when the validation score does not improve for 15 consecutive epochs. All experiments are conducted on NVIDIA RTX A6000 GPUs with PyTorch 2.1.1 and CUDA 11.8.

For video, we adopt dense clip sampling with a frame interval of four and an overlap ratio of 0.5 between consecutive clips, and the frame rate is normalized across recordings so that clip durations remain comparable between datasets. Only the last block of each pretrained video backbone is unfrozen, while the remaining blocks are kept frozen to preserve large-scale pretrained representations under the limited size of paired EEG-video data. In contrast to the video branch, all EEG backbones are fully fine-tuned, since EEG foundation models are pretrained on recording conditions that differ substantially from the affective stimuli considered here.

Learning rates and weight decay follow the configurations recommended for each backbone, as summarized in Table A6. The supervised video transformers are trained with a smaller learning rate of 2×10^{-5} and a weight decay of 0.02, whereas the masked-autoencoder backbones, all EEG backbones, and EVER use a learning rate of 1×10^{-4} and a weight decay of 0.05. EVER inherits the optimization setting of its paired backbones, so that the comparison against the unimodal models isolates the effect of the proposed fusion rather than differences in the optimization budget.

Multi-modal Baselines. Milmer (Wang et al., 2026) is a multi-modal architecture and is therefore evaluated as released. EAV (Lee et al., 2024), in contrast, is a dataset paper that provides modality-specific baselines but no joint fusion model. To obtain a faithful Video+EEG comparison, we pair its two published branches with a standard late-fusion head. The EEG branch follows the released ShallowConvNet (Schirrmeister et al., 2017) baseline: the EEG signal is downsampled to the 100 Hz target rate of EAV, band-pass filtered between 0.5 and 45 Hz, and segmented into five-second windows of 500 samples before the convolutional encoder. The video branch follows the per-frame ViT-B/16 (Dosovitskiy et al., 2021) protocol of EAV, in which every frame is classified independently and the frame-level logits are averaged into a sequence-level prediction. The two branches are combined by concatenating their logits and applying a single linear layer, which is the standard decision-level fusion used in the literature, and the model is trained under the subject-independent protocol as all other benchmark.

A.7.2 DETAILS OF DATASET

To further describe the details of the datasets, we visualize the 3D scatter plots of the self-assessment annotations for the Emognition (Saganowski et al., 2022) and MDMER (Yang et al., 2024) datasets as shown in Fig. A2. Each axis in the visualization corresponds to the respective annotation dimensions. The Emognition dataset is originally annotated along the dimensions of valence–arousal–motivation, while MDMER is based on valence–arousal–dominance. We observe that Emognition annotations are unevenly distributed across the space, whereas MDMER covers a wider range but remains concentrated near the middle scores. To address such imbalance and concentration issues during training, we adopt a combination of weighted cross-entropy loss and focal loss. We focus our prediction on valence and arousal, which are the most widely adopted dimensions in emotion research (*e.g.*, the circumplex model of affect) and serve as a common, interpretable basis for evaluating emotional states. Figure A3 and Figure A4 show representative 32-frame clips from the two datasets, illustrating that facial responses evolve gradually within a clip and motivating the clip-level temporal modeling of the VCG prior.



Figure A3: Sample of video frames on the Emognition dataset with disgust stimuli.



Figure A4: Sample of video frames on the MDMER dataset.