EXIST: EXPLICIT-GUIDED SPATIOTEMPORAL SUPERVISION FOR EEG EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion recognition in active emotional expression scenarios, such as clinical interviews, relies on audio and video, which provide rich affective cues, but cannot be collected from privacy-sensitive patients. EEG offers a privacy-preserving alternative, yet EEG-based emotion recognition falls behind audiovisual systems due to low signal-to-noise ratio and strong inter-subject variability. We propose **explicit-to-implicit supervision**, a distillation paradigm that leverages explicit modalities (video and audio) to supervise the implicit modality (EEG) during training and discards them at inference time. Current distillation methods cause negative transfer in this context because they align features or logits across incompatible spaces, and neglect the spatiotemporal structure of the EEG. Therefore, we propose **EXIST** (*EXplicit-guided Implicit SpatioTemporal supervision for EEG*), which performs *EEG-native* reasoning using audiovisual signals as structural supervision within the EEG manifold, enabling EEG-to-EEG transfer while avoiding modality conflict. In particular, we design two components: Cortical Dependency Distillation (CDD) and Emotional Dynamics Distillation (EDD). CDD transfers emotion-relevant cortical dependency priors inferred from audiovisual context into EEG-native spatial reasoning, whereas EDD focuses on learning emotionally salient temporal dynamics rather than enforcing direct imitation across modalities. Experiments on EAV and PME4 demonstrate that EXIST achieves state-of-the-art EEG-only emotion recognition under a subject-independent protocol, outperforming both prior EEG-only and existing cross-modal distillation methods.

1 INTRODUCTION

Automatic emotion recognition has attracted growing interest in affective computing. Two modalities are commonly employed. Explicit modalities such as audio and video expose behaviorally observable affective cues, while implicit modalities such as EEG reflect neural substrates of emotion without behavioral expression (Poria et al., 2017; Duan et al., 2013). These two families have largely evolved in parallel, with multimodal fusion (Zadeh et al., 2017) and EEG-based approaches (Zhong et al., 2020; Ding et al., 2025) developed independently. While most EEG-based works focus on controlled stimulus-elicitation settings (Koelstra et al., 2011; Zheng & Lu, 2015), active dialogue scenarios such as clinical interviews, mental health chatbots, and affective tutoring systems are more clinically common (Ringeval et al., 2019), in which spontaneous speech and facial expressions provide particularly rich affective cues. Yet the patients most in need of assessment, those with depression, social anxiety, or PTSD, are precisely the most likely to refuse audiovisual recording, creating a clear privacy barrier.

EEG mitigates this concern by avoiding directly revealing of identifiable facial and vocal cues, while being collected non-invasively via wearable headsets (Duan et al., 2013; Zheng & Lu, 2015). Yet, EEG-based recognition falls behind audiovisual systems (Fig. 1, left) (Muhammad et al., 2026), due to low signal-to-noise ratio, strong inter-subject variability (Ding et al., 2025; Jang et al., 2025), and the absence of surface behavioral cues. This motivates a new paradigm: rather than fusing the two modalities, we propose **explicit-to-implicit supervision**, which exploits the affective cues of explicit modalities (*e.g.* vision and audio) to supervise EEG learning at training and withdraws them at inference.

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.

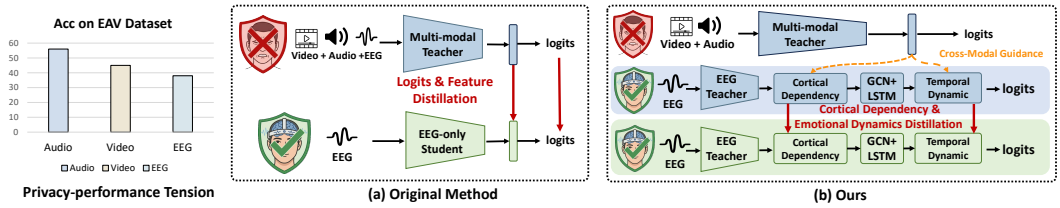


Figure 1: **Left**: Audio and video outperform EEG on EAV yet cannot be collected from privacy-sensitive patients; knowledge distillation resolves this by supervising EEG at training time and discarding explicit modalities at inference. **Right (a)** Conventional methods align features or logits across the modality gap, inducing negative transfer. **Right (b)** EXIST uses audiovisual context to guide the EEG teacher's cortical dependency prior to and emotional temporal dynamics, which is then distilled into the EEG student via intra-modal EEG-to-EEG transfer.

As shown in Fig. 1 right, knowledge distillation (Hinton et al., 2015) is a natural mechanism to achieve this, and prior work has explored distilling explicit modality teachers into EEG students (Zhang et al., 2022; Liu et al., 2023). However, these methods operate at the feature or logit level (Fig. 1 (a)), leaving two unsolved problems. First, EEG occupies a electrode-topology-constrained space (Zhong et al., 2020) that is incompatible with explicit-modality embedding spaces, and direct feature or logit alignment causes negative transfer (Mirzadeh et al., 2020; Cho & Hariharan, 2019). Second, existing EEG distillation methods treat EEG as a flat feature vector, discarding both the functional electrodes connectivity and the non-uniform temporal affective saliency (Song et al., 2021).

We propose **EXIST** (*EXplicit-guided Implicit SpatioTemporal supervision for EEG*), motivated by attention distillation (Zagoruyko & Komodakis, 2016). Rather than replicating teacher features, the student learns *where* in the cortex and *when* to attend during the expression. To achieve this, we design a teacher that consists of an explicit audiovisual stream and an implicit EEG stream (Fig. 1 (b)). Fused audiovisual representations (*i.e.*, explicit stream) serve as *affective resonance probes* that guide the EEG teacher's cortical dependency and emotional dynamics. Then, the EEG student learns to reproduce through intra-modal distillation. Since both the teacher and the student operate in the EEG space, and the audiovisual context serves only as a reference signal, the modality gap is resolved by design. In EXIST, we introduce Cortical Dependency Distillation (CDD) and Emotional Dynamics Distillation (EDD). CDD transfers audiovisual guidance into inter-electrode connectivity, identifying which cortical regions are most relevant to the expressed emotion. EDD transfers it into temporal attention, identifying which moments carry the most affective information. Together, CDD and EDD enable the EEG model to capture a richer spatiotemporal affective structure than it could be derived from EEG signals alone. Our main contributions are:

- We propose EXIST: an explicit-guided spatiotemporal supervision framework that uses auxiliary audiovisual signals as *affective resonance probes* to generate supervision targets natively within the EEG space, avoiding cross-modal feature alignment.
- We introduce Cortical Dependency Distillation (CDD) and Emotional Dynamics Distillation (EDD) to transfer cortical dependency priors and emotionally salient temporal dynamics from audiovisual context into EEG-native spatial reasoning.
- EXIST achieves state-of-the-art performance among all EEG-only and cross-modal distillation methods under a subject-independent protocol, validating the effectiveness of the proposed method on EAV and PME4.

2 Related Work

2.1 Multimodal and EEG-Based Emotion Recognition

Multimodal emotion recognition typically fuses speech, facial action units, text, and physiological signals (Poria et al., 2017; Tsai et al., 2019). Most benchmarks rely on controlled stimulus-elicitation settings (Koelstra et al., 2011; Yang et al., 2024), though active dialogue is more clinically relevant (Ringeval et al., 2019). On the EEG side, GNN-based methods have shown strong results by modeling electrode topology (Zhong et al., 2020; Song et al., 2021), and domain adaptation has

helped with inter-subject variability (Li et al., 2019). However, when EEG is used in multimodal systems, its spatiotemporal structure is largely ignored (Poria et al., 2017). A further limitation is that most methods assume that all modalities are available at inference time, which is impractical under privacy constraints.

2.2 Knowledge Distillation and Attention Transfer

Knowledge distillation (Hinton et al., 2015) transfers knowledge from teacher to student via soft label supervision. This idea has been extended to intermediate feature matching (Romero et al., 2015), relation-based (Park et al., 2019), and contrastive (Tian et al., 2020) objectives. In the EEG emotion recognition domain, EmotionKD (Liu et al., 2023) applies cross-modal distillation across physiological signals, but operates at the feature level and does not solve the modality gap. Attention transfer (Zagoruyko & Komodakis, 2016) distills attention maps rather than features or logits, letting the student learn *where* to attend without replicating the full representation of the teacher. This has proven effective in Transformer-based models, including self-attention distillation for task-agnostic compression (Wang et al., 2020) and vision transformers (Li et al., 2024). For EEG, however, attention transfer faces a structural problem. Attention maps from speech or video encoders are defined over their own feature spaces and do not map onto the EEG's electrode graph or temporal windows (Zhong et al., 2020). Rather than borrowing attention maps from explicit modalities, we compute attention targets directly in the EEG space as cross-modal affinities between audiovisual features and each electrode or temporal window. The student, therefore, remains aligned with EEG's own structure while still benefiting from explicit-modality guidance.

3 Methodology

Cross-modal distillation provides a privacy-preserving way to leverage audiovisual teachers for EEG emotion recognition, but aligning EEG representations with audiovisual features or logits causes negative transfer due to the modality gap. We propose EXIST, which resolves this through EEG-to-EEG knowledge transfer. Audiovisual context supervises EEG-native attention entirely within the EEG space via Cortical Dependency Distillation (CDD) and Emotional Dynamics Distillation (EDD), capturing which cortical regions and temporal moments of the EEG are most relevant to emotion.

3.1 EXIST: EXplicit-guided Implicit SpatioTemporal Supervision

As illustrated in Fig. 2, EXIST follows a two-stream teacher–student design. The explicit stream (Audio & Video Teacher) encodes audiovisual signals into an affective context $\mathbf{c}$, which guides the EEG teacher to produce AV-informed guidance along spatial and temporal dimensions: cortical dependency across electrodes and emotional dynamics across temporal windows. The EEG student then learns to reproduce these distributions through distillation, without any access to the context $\mathbf{c}$ at inference. Unlike conventional cross-modal distillation, which suffers from a modality gap, both teacher and student operate in the EEG space, and $\mathbf{c}$ serves only as a reference signal. We partition EEG signal into K non-overlapping temporal windows, each modeled as a spatial graph $\mathcal{G}^{(k)} = (\mathcal{V}, \mathbf{X}^{(k)}, \mathbf{A}^{(k)})$, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ is the electrode set, $\mathbf{X}^{(k)} \in \mathbb{R}^{N \times T_w}$ is the EEG signal, and $\mathbf{A}^{(k)} \in \mathbb{R}^{N \times N}$ is the inter-electrode adjacency.

Explicit stream (Audio & Video Teacher). Audio $\mathbf{a}$ and video $\mathbf{v}$ are encoded by modality-specific BiLSTMs with attention pooling, then fused via a linear projection with ReLU into an affective context $\mathbf{c} \in \mathbb{R}^d$, which serves as the affective resonance probe for CDD and EDD. The explicit stream predicts the emotion label $\hat{y}^{\text{AV}}$ from $\mathbf{c}$.

Implicit stream (EEG Teacher). The EEG teacher processes EEG with audiovisual guidance from $\mathbf{c}$ at both the cortical dependency and emotional dynamics steps. At the cortical dependency step, $\mathbf{c}$ acts as a query in a cross-attention operation over electrode features, producing per-electrode resonance scores that modulate the GCN adjacency $\mathbf{A}^{(k)}$ and yield the cortical dependency matrix $\mathbf{S}^{T,(k)}$. Node features are then aggregated by the GCN and converted into window summaries $\mathbf{s}^{(k)}$. At the emotional dynamics step, a BiLSTM encodes $\{\mathbf{s}^{(k)}\}$ into hidden states $\{\mathbf{h}^{(k)}\}$, and $\mathbf{c}$ again

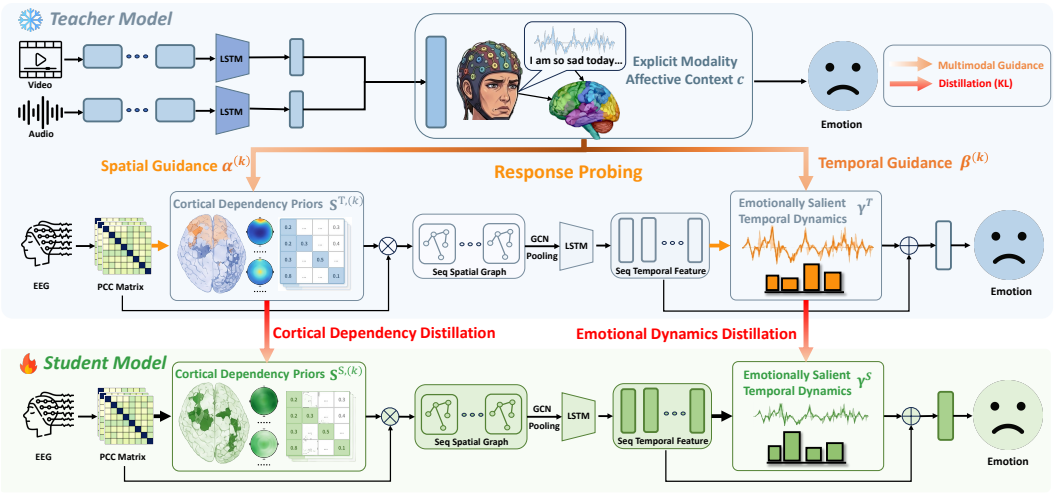


Figure 2: Overview of **EXIST**. The explicit stream encodes audio and video into affective context $\mathbf{c}$, which guides the EEG teacher to produce cortical dependency and emotional dynamics supervision via CDD and EDD. The EEG student learns to reproduce these supervision signals through KL distillation. At inference, the explicit stream and EEG teacher are discarded.

acts as a query over window hidden states to produce resonance scores that modulate the emotional dynamics weights $\boldsymbol{\gamma}^T \in \Delta^{K-1}$. The trial-level representation $\mathbf{z}^T = \sum_k \gamma_k^T \mathbf{h}^{(k)}$ is passed to a classifier to predict $\hat{y}^T$.

EEG student. The EEG student shares the same architecture as the teacher but has no access to $\mathbf{c}$. Its cortical dependency matrix $\mathbf{S}^{S,(k)}$ is derived from self-attention over electrode features, and its emotional dynamics weights $\boldsymbol{\gamma}^S$ are from self-attention over window hidden states. During training, KL distillation encourages the student to match the teacher's AV-guided distributions, $\mathbf{S}^{S,(k)} \rightarrow \mathbf{S}^{T,(k)}$ and $\boldsymbol{\gamma}^S \rightarrow \boldsymbol{\gamma}^T$, so that the student learns which cortical regions and which temporal moments are most relevant to emotion without access to explicit-modality signals.

3.2 Structural Dynamics Distillation

EEG exhibits rich spatial structure across electrodes and temporal dynamics across windows, yet existing distillation methods flatten it into vectors and ignore both. We address this with CDD and EDD, which use audiovisual context $\mathbf{c}$ to guide the cortical dependency across electrodes and emotional dynamics across windows, which the student then learns to reproduce through distillation.

3.2.1 Cortical Dependency Distillation (CDD)

CDD leverages cross-modal supervision to model functional connectivity across electrodes, capturing how teacher signals reshape spatial EEG activations to identify where in the cortex emotions are expressed. It implements cross-modal electrode resonance. For each temporal window k, it computes a resonance score for each electrode by measuring the affinity between the affective context $\mathbf{c}$ and the electrode's local EEG feature. These scores are added to the cortical dependency matrix $\mathbf{S}^{T,(k)}$ of the EEG teacher, steering inter-electrode connectivity toward emotion-relevant cortical regions. The student then learns to reproduce this connectivity pattern through KL distillation, without any cross-modal feature alignment. Fig. 3 (a) illustrates the detailed design of the CDD.

Student Cortical Dependency. For window k with node features $\mathbf{X}^{(k)} \in \mathbb{R}^{N \times F}$, the student computes a cortical dependency matrix via two linear projections and a scaled dot-product attention:

$$\mathbf{S}^{S,(k)} = \mathrm{softmax}_{\mathrm{row}}\left(\frac{\left(\mathbf{X}^{(k)}\mathbf{W}_1^\top\right)\left(\mathbf{X}^{(k)}\mathbf{W}_2^\top\right)^\top}{\tau}\right) \in \mathbb{R}^{N \times N}, \tag{1}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_s \times F}$ are learnable and $\tau > 0$ controls the sharpness of the distribution.

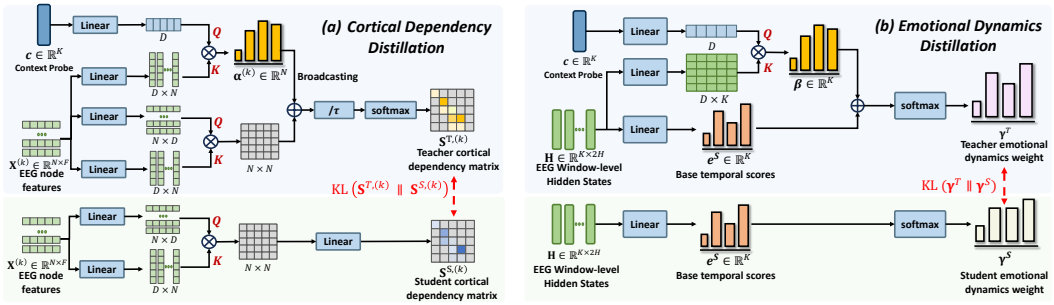


Figure 3: Details of CDD and EDD. CDD injects an AV-guided per-electrode resonance score into the cortical dependency matrix, directing inter-electrode connectivity toward emotion-relevant cortical regions. EDD injects an AV-guided per-window resonance score into the emotional dynamics weight, shifting the distribution toward emotionally salient temporal windows.

Teacher Cortical Dependency. The teacher augments the same dot-product with an AV-guided per-electrode resonance score. Using $\mathbf{c}$ as a query and each electrode's feature as a key:

$$\alpha_i^{(k)} = (\mathbf{W}_q\,\mathbf{c})^\top \left(\mathbf{W}'_k\,\mathbf{x}_i^{(k)}\right), \quad i = 1, \ldots, N, \tag{2}$$

where $\mathbf{W}_q \in \mathbb{R}^{d_s \times d}$ and $\mathbf{W}'_k \in \mathbb{R}^{d_s \times F}$ are learnable linear projections. Broadcasting $\boldsymbol{\alpha}^{(k)}$ as a column-wise offset yields teacher cortical dependency matrix:

$$\mathbf{S}^{T,(k)} = \mathrm{softmax}_{\mathrm{row}}\left(\frac{\left(\mathbf{X}^{(k)}\mathbf{W}_1^\top\right)\left(\mathbf{X}^{(k)}\mathbf{W}_2^\top\right)^\top + \mathbf{1}_N \boldsymbol{\alpha}^{(k)\top}}{\tau}\right) \in \mathbb{R}^{N\times N}. \tag{3}$$

$\alpha_i^{(k)}$ reflects how strongly electrode i resonates with $\mathbf{c}$. Adding $\alpha_i^{(k)}$ to the i-th column of $\mathbf{S}^{T,(k)}$ uniformly increases the attention that all other electrodes pay to electrode i, making it a more influential hub when it is affectively salient. This column-wise propagation directs the graph topology toward emotion-relevant cortical regions without modifying each pairwise relationship independently. The final adjacency combines the cortical dependency matrix with a Pearson Correlation Coefficient (PCC) prior $\mathbf{P} \in \mathbb{R}^{N\times N}$, whose construction is detailed in Appendix A:

$$\mathbf{A}^{M,(k)} = \mathbf{S}^{M,(k)} \odot \mathrm{ReLU}(\mathbf{P}) + \mathbf{I}, \qquad M \in \{T, S\}, \tag{4}$$

where $M = T$ and $M = S$ denote the teacher and student, respectively. Each adjacency matrix is degree-normalized to $\hat{\mathbf{A}}^{M,(k)}$ and passed through two GCN layers (Kipf & Welling, 2017). The resulting node representations are flattened across the N electrodes to obtain the window-level representation $\mathbf{s}^{(k)} \in \mathbb{R}^{Nd_{\mathrm{out}}}$.

CDD Loss. Both $\mathbf{S}^{T,(k)}$ and $\mathbf{S}^{S,(k)}$ are row-stochastic matrices, each row forming a probability distribution over the N electrodes. We align them by minimizing the average per-row KL divergence:

$$\mathcal{L}_{\mathrm{cdd}} = \frac{1}{BKN}\sum_{b=1}^{B}\sum_{k=1}^{K}\sum_{i=1}^{N} \mathrm{KL}\left(\mathbf{S}_{b,i,:}^{T,(k)} \,\middle\|\, \mathbf{S}_{b,i,:}^{S,(k)}\right), \tag{5}$$

where B is the batch size. This transfers cortical dependency supervision from teacher to student without any cross-modal feature alignment.

3.2.2 Emotional Dynamics Distillation (EDD)

EDD models emotional dynamics by reweighting the EEG teacher's attention across temporal windows under audiovisual guidance. For each window, $\mathbf{c}$ queries the window's EEG hidden state through cross-attention to compute a per-window resonance score. These scores are added to the EEG teacher's emotional dynamics weights, shifting the distribution toward windows most consistent with the audiovisual emotional content. The student then learns to reproduce this weighting through KL distillation, localizing *when* emotion is expressed within the EEG representational space. Fig. 3 (b) illustrates the detailed design.

Student Emotional Dynamics. The BiLSTM encodes $\{\mathbf{s}^{(k)}\}$ into hidden states $\mathbf{H} \in \mathbb{R}^{K\times 2H}$. The student scores each window via additive attention (Bahdanau et al., 2014):

$$e_k^S = \mathbf{v}^\top \tanh\Big(\mathbf{W}_t\, \mathbf{h}^{(k)}\Big), \qquad \boldsymbol{\gamma}^S = \mathrm{softmax}(\mathbf{e}^S) \in \mathbb{R}^K, \tag{6}$$

where $\mathbf{W}_t \in \mathbb{R}^{2H\times 2H}$ and $\mathbf{v} \in \mathbb{R}^{2H}$ are learnable. $\boldsymbol{\gamma}^S$ are the student's emotional dynamics weights, and the final EEG representation is $\mathbf{z} = \sum_k \gamma_k^S\, \mathbf{h}^{(k)}$.

Teacher Emotional Dynamics. The teacher augments the same additive score with an AV-guided window resonance bias computed via cross-attention between $\mathbf{c}$ and each window's hidden state:

$$\beta_k = (\mathbf{W}_q\, \mathbf{c})^\top \Big(\mathbf{W}_k''\, \mathbf{h}^{(k)}\Big), \qquad \boldsymbol{\gamma}^T = \mathrm{softmax}(\mathbf{e}^T + \boldsymbol{\beta}) \in \mathbb{R}^K, \tag{7}$$

where $\mathbf{e}^T$ denotes the teacher's own temporal scores, computed in the same manner as the student's temporal scores $\mathbf{e}^S$ in Eq. equation 6; $\mathbf{W}_q \in \mathbb{R}^{d_s\times d}$ and $\mathbf{W}_k'' \in \mathbb{R}^{d_s\times 2H}$ are learnable linear projections. $\boldsymbol{\gamma}^T$ are the teacher's emotional dynamics weights. β_k shifts their distribution toward windows whose EEG activity is most consistent with the audiovisual emotional content.

EDD Loss. Both $\boldsymbol{\gamma}^T$ and $\boldsymbol{\gamma}^S$ are K-dimensional probability distributions over the same temporal windows. We align them by minimizing their KL divergence:

$$\mathcal{L}_{\text{edd}} = \mathrm{KL}\Big(\boldsymbol{\gamma}^T \,\Big\|\, \boldsymbol{\gamma}^S\Big). \tag{8}$$

This encourages the student to assign higher weights to the most emotionally informative windows.

3.3 Training Objective

First, the explicit stream and EEG teacher are jointly trained:

$$\mathcal{L}_{\text{T}} = \mathcal{L}_{\text{cls}}^{\text{AV}} + \mathcal{L}_{\text{cls}}^{T}, \tag{9}$$

ensuring $\mathbf{c}$ captures genuine affective content and the EEG teacher produces meaningful cortical dependency and emotional dynamics distributions for distillation. Then, the teacher is frozen, and the EEG student is trained with:

$$\mathcal{L}_{\text{S}} = \mathcal{L}_{\text{cls}}^{S} + \lambda_{\text{cdd}}\, \mathcal{L}_{\text{cdd}} + \lambda_{\text{edd}}\, \mathcal{L}_{\text{edd}}, \tag{10}$$

where $\mathcal{L}_{\text{cls}}^S$ is the student classification loss, and $\lambda_{\text{cdd}}, \lambda_{\text{edd}} > 0$ balance the two distillation terms.

4 Experiments

4.1 Datasets and Experimental Setup

We evaluate EXIST on two active-speaking emotion datasets. The EAV dataset (Lee et al., 2024) contains multimodal recordings from 42 participants across five emotion categories (*Neutral*, *Anger*, *Happiness*, *Sadness*, and *Calmness*), with audio, video, and EEG captured simultaneously. Speaking segments are extracted and partitioned into non-overlapping 5-second windows as trial-level inputs. The PME4 dataset (Chen et al., 2022) contains synchronized audio and EEG recordings from 11 participants under the same active speaking condition. We select four emotion categories (*Happy*, *Anger*, *Fear*, and *Sad*) for evaluation. EEG signals are preprocessed into differential entropy (DE) features (Duan et al., 2013) across five frequency bands as node-level inputs to the graph model. We perform three random subject-independent splits with an approximately 6:2:2 training/validation/test ratio and three training seeds per split, while keeping all samples from each subject within the same split. We report the mean and standard deviation of accuracy (Acc) and weighted F1-score (F1) over the resulting nine runs. All experiments are implemented in PyTorch and trained on an NVIDIA GeForce RTX 5090 GPU using the Adam optimizer with a learning rate of 5×10^{-4} and a batch size of 64.

Table 1: Comparison with EEG-only baselines on EAV and PME4 under the cross-subject setting.† indicates knowledge distillation from explicit modalities during training. Best results are in bold.

Method	Venue	EAV		PME4	
		Acc	F1	Acc	F1
EEGNet (Lawhern et al., 2018)	JNE'18	38.41 ± 2.59	35.16 ± 4.22	27.15 ± 0.35	26.82 ± 0.21
DGCNN (Song et al., 2018)	TAFFC'18	39.83 ± 2.37	38.40 ± 2.18	24.18 ± 0.43	15.90 ± 4.59
RGNN (Zhong et al., 2020)	TAFFC'20	31.57 ± 3.96	24.95 ± 5.58	23.15 ± 2.35	15.05 ± 2.12
EEGFormer (Wan et al., 2023)	FN'23	38.08 ± 0.48	35.58 ± 0.82	25.15 ± 0.15	16.90 ± 1.70
LaBraM (Jiang et al., 2024)	ICLR'24	30.76 ± 0.57	30.98 ± 0.71	24.38 ± 1.62	16.21 ± 1.01
iTransformer (Liu et al., 2024)	ICLR'24	30.81 ± 0.69	28.04 ± 0.29	25.78 ± 0.28	20.94 ± 2.98
DMMR (Wang et al., 2024)	AAAI'24	39.29 ± 2.69	38.72 ± 2.29	23.66 ± 0.41	20.31 ± 0.26
EMT (Ding et al., 2025)	TNNLS'25	35.67 ± 1.12	33.92 ± 0.79	28.39 ± 0.39	27.05 ± 1.94
STRFL (Hu et al., 2025)	TAFFC'25	31.23 ± 0.29	24.22 ± 2.11	25.59 ± 0.09	16.67 ± 0.34
EEG-SCMM (Liu et al., 2025)	ACM MM'25	39.52 ± 1.11	38.77 ± 1.03	29.03 ± 0.67	26.83 ± 0.68
Ours†	–	**41.49** ± 2.33	**40.04** ± 1.42	**31.03** ± 0.29	**29.93** ± 0.26

Table 2: Comparison with conventional knowledge distillation methods under the same teacher–student backbones, cross-subject protocol, and EEG-only inference. Best results are in bold.

Method	Venue	EAV		PME4	
		Acc	F1	Acc	F1
Student w/o KD	–	38.72 ± 2.58	38.04 ± 2.55	27.76 ± 0.24	26.81 ± 0.33
KD (Hinton et al., 2015)	NeurIPS'14	37.95 ± 3.22↓0.77	35.57 ± 2.47↓2.47	28.41 ± 0.16↑0.65	23.47 ± 3.79↓3.34
FitNets (Romero et al., 2015)	ICLR'15	38.67 ± 2.06↓0.05	37.02 ± 0.66↓1.02	28.66 ± 1.59↑0.90	22.27 ± 1.27↓4.54
NST (Huang & Wang, 2017)	ArXiv'17	36.98 ± 3.99↓1.74	34.75 ± 4.68↓3.29	26.40 ± 2.60↓1.36	21.99 ± 0.15↓4.82
EmotionKD (Liu et al., 2023)	ACM MM'23	39.78 ± 1.41↑1.06	39.34 ± 1.42↑1.30	27.16 ± 0.66↓0.60	26.91 ± 1.24↑0.10
CDGKD (Wu et al., 2025)	IJCAI'25	38.27 ± 2.39↓0.45	38.84 ± 2.18↑0.80	25.53 ± 1.28↓2.23	22.28 ± 2.79↓4.53
AMBOKD (Li et al., 2025)	TNNLS'25	38.35 ± 1.97↓0.37	37.89 ± 1.75↓0.15	27.28 ± 1.22↓0.48	21.30 ± 1.35↓5.51
CMCRD (Kan et al., 2026)	TAFFC'26	39.53 ± 2.12↑0.81	39.06 ± 1.77↑1.02	24.91 ± 0.09↓2.85	12.31 ± 2.45↓14.50
Ours	–	**41.49** ± 2.33↑2.77	**40.04** ± 1.42↑2.00	**31.03** ± 0.29↑3.27	**29.93** ± 0.26↑3.12

4.2 Overall Evaluation

Comparison with EEG-only Baselines. Table 1 compares EXIST with representative EEG-only methods under the cross-subject protocol. On EAV, EXIST outperforms all baselines, surpassing the strongest competitor DGCNN by 1.66% in accuracy. On PME4, most baselines perform near chance level, likely due to the limited number of EEG electrodes (8 channels) and the resulting low signal quality, which makes cross-subject generalization particularly challenging. EXIST still achieves the best performance on PME4, demonstrating that explicit-modality guidance provides discriminative structure even under degraded EEG conditions.

Comparison with Knowledge Distillation Baselines. Table 2 compares EXIST with conventional distillation methods under the same teacher and student backbone, including general methods KD (Hinton et al., 2015), FitNets (Romero et al., 2015), and NST (Huang & Wang, 2017), and physiological signal methods EmotionKD (Liu et al., 2023), CDGKD (Wu et al., 2025), AMBOKD (Li et al., 2025), and CMCRD (Kan et al., 2026). Many conventional distillation methods exhibit negative transfer, especially on PME4, where most methods degrade one or both metrics relative to the no-distillation baseline. Although several methods provide modest improvements on individual metrics, none consistently improves both accuracy and F1 across the two datasets. In contrast, EXIST consistently improves the baseline, gaining +2.77% accuracy and +2.00% F1 on EAV, and +3.27% accuracy and +3.12% F1 on PME4. This confirms that effective distillation in this setting benefits from supervision signals defined natively within the EEG space, including cortical dependency and emotional dynamics, rather than conventional feature- or logit-level targets.

4.3 Ablation Study and Analysis

Contribution of CDD and EDD. Table 3 isolates the contribution of each distillation module on EAV. CDD alone yields +2.31% in accuracy and +1.95% in F1, confirming that directing inter-electrode connectivity toward emotion-relevant cortical regions provides discriminative structure beyond what EEG encodes by itself. EDD alone contributes a comparable +1.82% in accuracy and +1.83% in F1, showing that localizing emotionally relevant temporal segments is equally important. Combining both modules achieves the best result, +2.77% in accuracy and +2.00% in F1. This demonstrates that CDD and EDD are complementary. CDD answers *where* in the cortex and EDD answers *when* the representation reflects emotions from EEG data. Together they capture the full spatiotemporal structure that explicit-modality guidance reveals.

Table 3: Ablation on CDD and EDD on EAV.

CDD	EDD	Acc (%)	F1 (%)
		38.72 ± 2.58	38.04 ± 2.55
✓		41.03 ± 3.03	39.99 ± 2.70
	✓	40.54 ± 1.80	39.87 ± 1.46
✓	✓	**41.49** ± 2.33	**40.04** ± 1.42

Effect of Distillation Supervision Target. Table 4 examines two design choices: the teacher source and the distillation target. Switching from a multimodal teacher to an AV-guided EEG teacher consistently improves all three baseline methods, confirming that reducing the modality gap at the teacher side reduces negative transfer. However, even with the EEG teacher, all three methods remain close to or marginally above the no-distillation baseline, showing that feature- and logit-level targets are insufficient even when the modality gap is partially addressed. EXIST addresses this by supervising cortical dependency and emotional dynamics natively within the EEG space, achieving 41.49% accuracy and 40.04% F1, surpassing the strongest EEG-teacher baseline by 1.80% accuracy and 2.19% F1.

Table 4: Effect of distillation supervision target on EAV. MM = Multimodal teacher, EEG-T = EEG teacher.

Method	Teacher	Acc (%)	F1 (%)
EEG student	–	38.72 ± 2.58	38.04 ± 2.55
KD	MM	37.95 ± 3.22	35.57 ± 2.47
	EEG-T	38.45 ± 2.79	36.96 ± 2.67
FitNets	MM	38.67 ± 2.06	37.02 ± 0.66
	EEG-T	39.69 ± 2.45	37.85 ± 2.05
NST	MM	36.98 ± 3.99	34.75 ± 4.68
	EEG-T	38.55 ± 2.11	37.60 ± 1.76
Ours	EEG-T	**41.49** ± 2.33	**40.04** ± 1.42

Effect of Explicit Guidance. Table 5 reports the effect of AV guidance on the EEG teacher. Without any guidance, the EEG teacher reduces to a standard EEG-only model, achieving 38.72% acc on EAV and 27.76% on PME4. Using spatial and temporal guidance bring the EEG teacher to 50.33% on EAV and 35.51% on PME4, closing much of the gap to the AV teacher upper bound. This confirms that $\mathbf{c}$ effectively steers EEG attention toward emotion-relevant structure that EEG signals alone cannot identify.

Table 5: Effect of explicit guidance on EEG representation quality. † indicates audio and visual modalities are used at inference.

Method	EAV		PME4	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)
AV teacher† (upper bound)	53.59 ± 3.05	53.01 ± 3.93	37.43 ± 2.57	33.60 ± 0.57
EEG teacher w/o guidance	38.72 ± 2.58	38.04 ± 2.55	27.76 ± 0.24	26.81 ± 0.33
EEG teacher w/ guidance†	50.33 ± 2.65↑11.61	48.91 ± 4.27↑10.87	35.51 ± 3.99↑7.75	32.58 ± 2.01↑5.77

Hyperparameter Analysis. Fig. 4 reports the sensitivity of EXIST to three key hyperparameters on EAV. λ_{cdd} peaks at 2.0, where moderate CDD supervision is beneficial but overly large weights suppress the student's own cortical dependency learning. λ_{edd} is optimal at 0.1. The attention temperature τ is best at 0.1, producing sharp distributions that concentrate on emotion-relevant electrodes; larger values reduce discriminability.

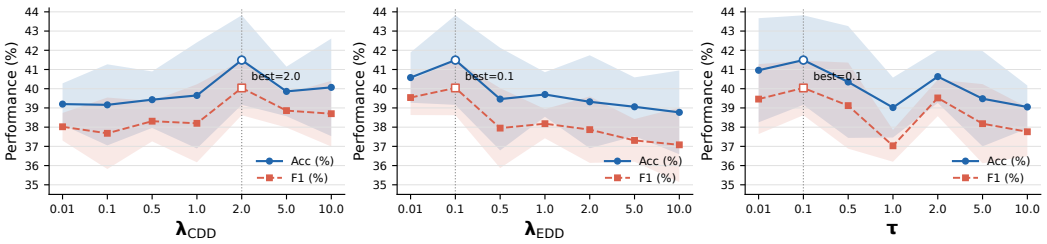
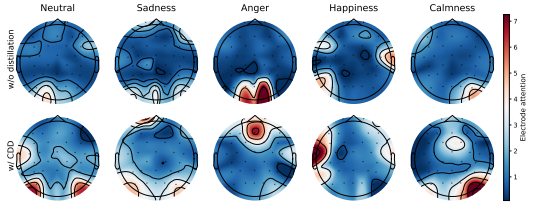
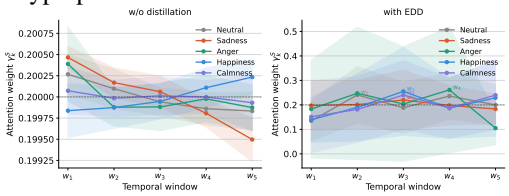


Figure 4: Sensitivity analysis of hyperparameters on EAV.



(a) Scalp topographic maps across five emotion categories, without (top) and with(bottom) CDD.



(b) Emotional dynamic weight $\boldsymbol{\gamma}^S$ with (right) and without (left) EDD. Without distillation, all curves cluster near uniform ($1/K = 0.2$); with distillation, emotion-specific patterns emerge with larger variance.

Figure 5: Interpretability analysis of CDD and EDD.

Scalp Topographic Analysis. Fig. 5a compares cortical dependency patterns before and after CDD distillation. Without distillation, electrode attention is sparse and relatively uniform across emotions. After distillation, more emotion-specific attention patterns emerge: neutral shows higher attention over bilateral occipital electrodes; sadness shows higher attention over the prefrontal and bilateral temporal–occipital regions; anger shifts attention toward the prefrontal region; happiness shows higher attention over the left temporal region; and calmness emphasizes the right temporal–occipital region. Overall, CDD distillation raises attention weights on emotion-relevant cortical regions, producing more discriminative cortical dependency representations.

Emotional Dynamics Analysis. Fig. 5b compares the student's emotional dynamics weights $\boldsymbol{\gamma}^S$ before and after EDD distillation. Without distillation, the weights of all five emotions remain close to the uniform baseline of 0.2, indicating that the student assigns equal importance to all temporal windows. After EDD distillation, the distributions diverge. Anger starts low at w_1, rises through w_2–w_3, then drops sharply at w_5. Happiness starts low at w_1 and peaks at w_3. Calmness rises gradually from w_1 to w_3 and then maintains a stable level. Neutral and Sadness remain stable across windows. More importantly, within-category variance expands substantially after distillation: different subjects express the same emotion at different moments, and EDD enables the student to attend to these person-specific affective peaks.

5 Conclusion

We present **EXIST**, an explicit-guided spatiotemporal supervision framework for EEG-based emotion recognition in active emotional expression scenarios. Audio and video carry rich affective cues but raise privacy concerns at inference. EXIST addresses this through a two-stream teacher design, where an audiovisual stream generates affective context that guides an EEG teacher stream via cross-attention, producing cortical dependency and emotional dynamics supervision natively within the EEG space. The EEG student learns through intra-modal distillation via CDD and EDD, requiring no cross-modal feature alignment. CDD supervises inter-electrode cortical dependency to identify which brain regions are most relevant to emotion, while EDD supervises emotional dynamics weights to identify which temporal moments carry the most affective information. Experiments on EAV and PME4 under a subject-independent protocol show state-of-the-art performance among EEG-only methods,and the interpretability analyses provide insights into the spatial and temporal decision patterns learned by the model. We hope this work encourages further exploration of explicit-to-implicit supervision as a general strategy for privacy-preserving affective computing.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with language polishing and improving the clarity and readability of the manuscript. We did not use generative AI tools for research idea generation, methodology design, experimental design, data analysis, code generation, or result interpretation. All AI-assisted text was carefully reviewed and revised by the authors. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing datasets and does not involve the collection of new human-subject data. All data are used in accordance with their original research protocols and licenses. Since the task involves EEG, audio, and visual signals, we recognize potential privacy concerns in affective computing. Our method requires audiovisual information only during training and uses EEG alone at inference.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training procedure, and distillation objectives in the main paper. Dataset preprocessing, input representations, temporal window settings, and baseline implementation details are provided in the appendix. Additional implementation details, hyperparameter settings, and cross-subject evaluation protocols are also documented to facilitate reproduction of the reported results. We will release the source code and configuration files upon publication.

REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Jin Chen, Tony Ro, and Zhigang Zhu. Emotion recognition with audio, video, eeg, and emg: a dataset and baseline approaches. *IEEE Access*, 10:13229–13242, 2022.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

Yi Ding, Chengxuan Tong, Shuailei Zhang, Muyun Jiang, Yong Li, Kevin JunLiang Lim, and Cuntai Guan. Emt: A novel transformer for generalized cross-subject eeg emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pp. 81–84. IEEE, 2013.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Fo Hu, Kailun He, Can Wang, Qinxu Zheng, Bin Zhou, Gang Li, and Yu Sun. Strflnet: Spatio-temporal representation fusion learning network for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*, 2025. doi: 10.1109/TAFFC.2025.3611173. URL https://ieeexplore.ieee.org/document/11168179/.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

Hyo-Jeong Jang, Hye-Bin Shin, and Seong-Whan Lee. Uncertainty-aware cross-modal knowledge distillation with prototype learning for multimodal brain-computer interfaces. *arXiv preprint arXiv:2507.13092*, 2025.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data (labram). In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Siyuan Kan, Huanyu Wu, Zhenyao Cui, Fan Huang, Xiaolong Xu, and Dongrui Wu. Cmcrd: Cross-modal contrastive representation distillation for emotion recognition. *IEEE Transactions on Affective Computing*, 2026.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 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.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c. URL http://arxiv.org/abs/1611.08024.

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.

Alexander C. Li, Yuandong Tian, Beidi Chen, Deepak Pathak, and Xinlei Chen. On the surprising effectiveness of attention transfer for vision transformers. In *Advances in Neural Information Processing Systems 37*, 2024.

Jinpeng Li, Shuang Qiu, Yuan-Yuan Shen, Cheng-Lin Liu, and Huiguang He. Multisource transfer learning for cross-subject eeg emotion recognition. *IEEE transactions on cybernetics*, 50(7): 3281–3293, 2019.

Zixing Li, Chao Yan, Zhen Lan, Xiaojia Xiang, Han Zhou, Jun Lai, and Dengqing Tang. Adaptive modality balanced online knowledge distillation for brain–eye–computer-based dim object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Qile Liu, Weishan Ye, Lingli Zhang, and Zhen Liang. Eeg-scmm: Soft contrastive masked modeling for cross-corpus eeg-based emotion recognition. arXiv:2408.09186, 2025. URL http://arxiv.org/abs/2408.09186.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Yucheng Liu, Ziyu Jia, and Haichao Wang. Emotionkd: a cross-modal knowledge distillation framework for emotion recognition based on physiological signals. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6122–6131, 2023.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Ghulam Muhammad, Sumayah Almuntasheri, Fadia Alenezi, Nwraan Alhadi, and Victor CM Leung. Eeg-based multimodal emotion recognition: Recent progress, challenges, and future directions. *ACM Transactions on Multimedia Computing, Communications and Applications*, 22(2): 1–28, 2026.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3967–3976, 2019.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.

Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pp. 3–12, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015.

Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3): 532–541, 2018.

Tengfei Song, Wenming Zheng, Suyuan Liu, Yuan Zong, Zhen Cui, and Yang Li. Graph-embedded convolutional neural network for image-based eeg emotion recognition. *IEEE Transactions on Emerging Topics in Computing*, 10(3):1399–1413, 2021.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Zhijiang Wan, Manyu Li, Shichang Liu, Jiajin Huang, Hai Tan, and Wenfeng Duan. Eegformer: A transformer–based brain activity classification method using eeg signal. *Frontiers in neuroscience*, 17:1148855, 2023.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33*, pp. 5776–5788. Curran Associates, Inc., 2020.

Yiming Wang, Bin Zhang, and Yujiao Tang. Dmmr: Cross-subject domain generalization for eeg-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 628–636, 2024.

Shuang Wu, Heng Liang, Yong Zhang, Yanlin Chen, and Ziyu Jia. A cross-modal densely guided knowledge distillation based on modality rebalancing strategy for enhanced unimodal emotion recognition. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16–22, 2025*, pp. 4236–4244, 2025.

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.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 1103–1114, 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Su Zhang, Chuangao Tang, and Cuntai Guan. Visual-to-eeg cross-modal knowledge distillation for continuous emotion recognition. *Pattern Recognition*, 130:108833, 2022.

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.

Peixiang Zhong, Di Wang, and Chunyan Miao. Eeg-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, 13(3):1290–1301, 2020.

A Appendix

Model Architecture Details

Table A.1 summarizes the architecture of the Teacher and Student models. The two share the same EEG backbone—SGCN followed by ABiLSTM—but differ in their attention modules: the Teacher injects $\mathbf{c}$ into both the cortical dependency and emotional dynamics steps via CDD and EDD, while the Student computes cortical dependency and emotional dynamics weights from EEG signals alone.

Table A.1: Teacher and Student architecture. † indicates Teacher-only components.

Module	Layer	Input dim	Output dim	Both / Teacher only
AV Encoder†	Audio BiLSTM + Attn Pooling	25	128	Teacher only
	Vision BiLSTM + Attn Pooling	161	128	Teacher only
	Fusion (concat)	256	256	Teacher only
SGCN (CDD)	Cortical Dependency ($\mathbf{W}_1$, $\mathbf{W}_2$)	5	5	Both
	AV Cross-Attn ($\mathbf{W}_q$, $\mathbf{W}'_k$)†	256, 5	32	Teacher only
	GCN Layer 1 + LayerNorm	5	64	Both
	GCN Layer 2 + LayerNorm	64	64	Both
ABiLSTM (EDD)	BiLSTM	1920	128	Both
	Emotional Dynamics (FC + $\mathbf{v}^\top$)	128	1	Both
	AV Cross-Attn ($\mathbf{W}_q$, $\mathbf{W}''_k$)†	256, 128	32	Teacher only
Classifier	Linear + BN + ReLU	128	64	Both
	Linear	64	5	Both

Pearson Correlation Coefficient (PCC) Prior

For each sample n and window t, we compute a $V \times V$ PCC matrix over the DE feature vectors of all electrode pairs. For electrodes i and j:

$$\mathbf{P}_{ij}^{(n,t)} = \frac{\sum_{f=1}^{F} \left(\mathbf{x}_{i,f}^{(n,t)} - \bar{\mathbf{x}}_i^{(n,t)}\right)\left(\mathbf{x}_{j,f}^{(n,t)} - \bar{\mathbf{x}}_j^{(n,t)}\right)}{\sqrt{\sum_{f=1}^{F} \left(\mathbf{x}_{i,f}^{(n,t)} - \bar{\mathbf{x}}_i^{(n,t)}\right)^2}\sqrt{\sum_{f=1}^{F} \left(\mathbf{x}_{j,f}^{(n,t)} - \bar{\mathbf{x}}_j^{(n,t)}\right)^2}} \tag{11}$$

where $\mathbf{x}_i^{(n,t)} \in \mathbb{R}^F$ is the DE feature vector of electrode i at window t, $F = 5$ is the number of frequency bands, and $\bar{\mathbf{x}}_i^{(n,t)}$ denotes its mean. The resulting matrix $\mathbf{P}^{(n,t)} \in [-1, 1]^{V \times V}$ encodes inter-electrode functional connectivity and serves as a static structural prior in the adjacency matrix (Eq. 4).

Dataset Details

EAV. EAV contains recordings from 42 subjects (aged 20–30) in a cue-based conversational setting, where subjects alternated between listening and speaking phases. Five emotion categories are covered: Neutral, Anger, Happiness, Sadness, and Calmness. We use Speaking trials only, segmented into 5s windows. EEG was recorded with a 30-channel BrainAmp system (Ag/AgCl electrodes) at 500 Hz, downsampled to 100 Hz, and bandpass filtered at 0.5–50 Hz. Audio and visual features are extracted with OpenSMILE (25-dim) and OpenFace (161-dim), respectively. We perform three random subject-level splits with 25/8/9 subjects for training, validation, and testing, respectively.

PME4. PME4 contains recordings from 11 subjects with acting experience. Each subject read a fixed neutral sentence while expressing one of 7 emotions across 350 trials. We use the 4 emotions:

Table A.2: Summary of datasets.

	EAV	PME4
Subjects	42	11
Classes	5	4
Scenario	Active dialogue	Posed expression
EEG Device	BrainAmp	–
EEG Ch.	30	8
SR (Hz)	100	1,000
Audio	OpenSMILE (25-dim)	OpenSMILE (25-dim)
Vision	OpenFace (161-dim)	–
Train samples	10,000	1,385
Val samples	3,200	399
Test samples	3,600	399

Table A.3: Input configurations and temporal settings of baseline methods. Time window indicates whether the input samples are explicitly constructed using temporal segmentation.

Method	Input Used in Our Experiments	Time Window
EEGNet	Raw EEG	None
DGCNN	DE features	1s non-overlapping window
EEGFormer	Raw EEG	None
LaBraM	Raw EEG	None
iTransformer	Raw EEG	None
EmT	rPSD	1s non-overlapping window
STRFL	Raw EEG	1s non-overlapping window
EEG-SCMM	DE features	None
RGNN	DE features	None
DMMR	Raw EEG	None
EXIST	DE features	1s non-overlapping window

Anger, Happiness, Sadness, and Fear. EEG was recorded at 5 kHz from 8 scalp electrodes (F3, Fz, F4, Cz, P3, Pz, P4, O2), downsampled to 1 kHz and bandpass filtered at 0.1–30 Hz. Audio features are extracted with OpenSMILE (25-dim). We perform three random subject-level splits with 7/2/2 subjects for training, validation, and testing, respectively. Table A.2 summarizes both datasets.

A.1 Baseline Implementation Details

To ensure a fair comparison, we preserve the original input design and model characteristics of each EEG-Only baseline method. All methods are evaluated under the same cross-subject protocol, and hyperparameters are selected using validation subjects only. For baseline reproduction, we follow the original input formulations of each method whenever possible. As summarized in Table A.3, EEGNet, EEGFormer, LaBraM, and iTransformer directly process EEG signals without explicit temporal window partitioning. DGCNN, EmT, and STRFL require temporal-window inputs, and therefore we use the same 1-second non-overlapping window setting as EXIST. Feature-based methods, including EEG-SCMM, RGNN, and DMMR, are evaluated using their corresponding extracted feature representations without introducing additional temporal segmentation. For LaBraM, we initialize the model using its original pretrained weights and fine-tune it separately on each dataset.

A.2 Deployment Setting

We consider an asymmetric training–deployment setting in which audiovisual signals are available only as privileged information during model development. Specifically, synchronized EEG and explicit-modality signals are collected from a consenting development cohort during training, whereas deployment users provide EEG only, with no audio or video required at inference. Thus,

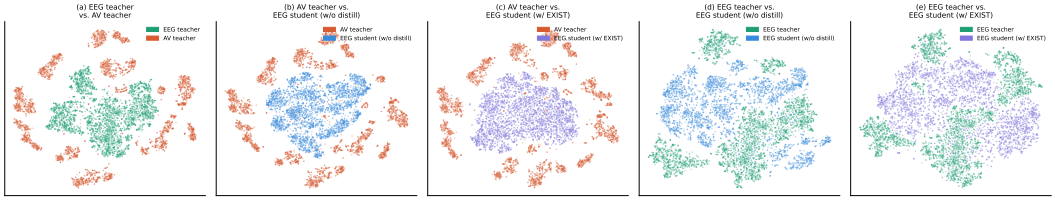


Figure A.1: t-SNE visualization of feature representations. *(a)* illustrates a visible separation between EEG and AV representations, consistent with the modality-gap hypothesis. *(b)(c)* AV teacher remains separated from both EEG student variants, showing that EXIST does not force EEG into the AV space. *(d)(e)* Without distillation, EEG teacher and student show limited overlap; with EXIST, their distributions become more overlapping despite no explicit feature regression, suggesting that CDD and EDD implicitly align the two EEG representations.

EXIST does not eliminate the need for audiovisual data during training; instead, it transfers knowledge from these training-only modalities into an EEG-only model for audiovisual-free deployment.

More Ablation Studies

Effect of α Injection. Table A.4 compares three ways of injecting the per-electrode resonance score α into the cortical dependency logits. Row-wise injection increases the attention that AV-salient electrodes pay others, treating them as active information senders. Element-wise injection strengthens connections only between pairs of mutually significant electrodes, requiring bilateral resonance. Column-wise injection (EXIST) increases the attention that all electrodes pay toward AV-salient electrodes, treating them as important information receivers. Column-wise achieves the best result, confirming that directing information flow toward emotion-relevant cortical regions is more effective than the alternatives.

Table A.4: Effect of α injection on EAV.

α injection	Acc (%)	F1 (%)
Row-wise	40.78 $\pm$ 1.68	39.93 $\pm$ 1.50
Element-wise	40.83 $\pm$ 2.44	39.62 $\pm$ 2.51
Column-wise (Ours)	**41.49** $\pm$ 2.33	**40.04** $\pm$ 1.42

Table A.5: Effect of the PCC prior on EAV.

Adjacency	Acc (%)	F1 (%)
S only	40.77 $\pm$ 2.08	38.82 $\pm$ 2.06
S + PCC	**41.49** $\pm$ 2.33	**40.04** $\pm$ 1.42

Effect of PCC Prior We further investigate the effect of incorporating the Pearson correlation coefficient (PCC) prior into the EEG adjacency matrix. As shown in Table A.5, using learned spatial dependency **S** alone achieves 40.77% accuracy and 38.82% F1. Incorporating the PCC prior improves the performance to 41.49% accuracy and 40.04% F1, corresponding to gains of 0.72% and 1.22%, respectively. These results show that the functional connectivity prior provides complementary structural information to the learned spatial dependency.

Feature Space Analysis. Fig. A.1 visualizes t-SNE projections of representations from the AV teacher, EEG teacher, and EEG student. Fig. (a) confirms that EEG and AV representations occupy fundamentally different feature spaces, explaining why direct cross-modal feature alignment induces negative transfer. Figs. (b) and (c) show that the AV teacher remains separated from both EEG student variants, confirming that EXIST does not force EEG representations into the AV space, consistent with its design of treating audiovisual context as a reference signal rather than a regression target. Comparing Figs. (d) and (e), without distillation the EEG teacher and student distributions show limited overlap; after EXIST distillation, their distributions exhibit greater overlap despite no explicit feature-level regression, suggesting that CDD and EDD implicitly encourage the student

to develop cortical dependency and emotional dynamics representations more consistent with the AV-guided EEG teacher.

Effect of AV Guidance on EEG Teacher. Table A.6 ablates the contribution of spatial and temporal AV guidance to the EEG teacher. Without any guidance, the teacher reduces to a standard EEG-only model at 38.72%. Removing spatial guidance (temporal only) raises accuracy to 45.04%, and removing temporal guidance (spatial only) reaches 46.06%—both substantially above the no-guidance baseline, confirming that each component independently contributes to the teacher's performance. Combining both achieves 50.33%, with spatial and temporal guidance providing complementary gains. These results confirm that AV-guided cortical dependency and emotional dynamics supervision substantially improves the EEG teacher, which in turn provides richer distillation targets for the student.

Table A.6: Effect of AV guidance on EEG teacher on EAV.

EEG Teacher Config	Acc (%)	F1 (%)
w/o guidance	38.72 ± 2.58	38.04 ± 2.55
w/o spatial	45.04 ± 1.96	43.61 ± 2.81
w/o temporal	46.06 ± 1.27	45.37 ± 1.50
w/ boths	**50.33** ± 2.65	**48.91** ± 4.27

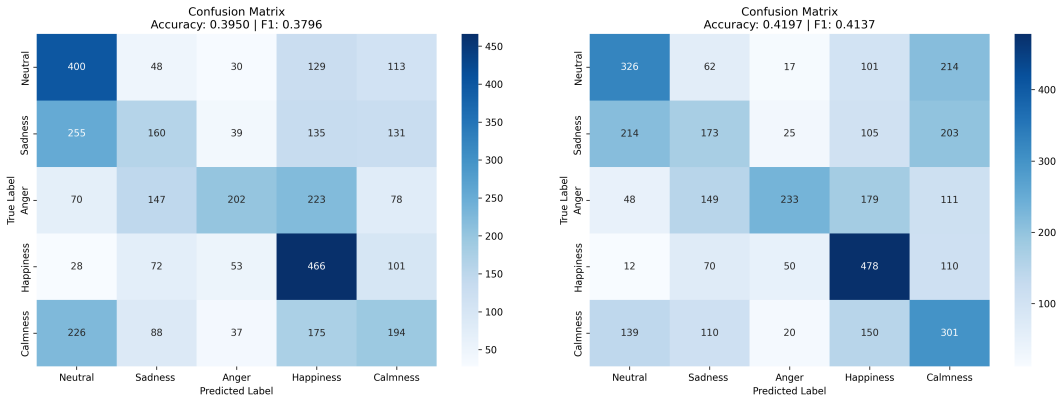
Table A.7: Effect of temporal window size on EAV. T denotes the number of windows per trial.

Window (s)	T	EEG Student		EEG Teacher	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)
0.1	50	38.67 ± 1.32	36.12 ± 1.41	43.90 ± 5.12	40.42 ± 7.53
0.5	10	40.50 ± 2.18	39.85 ± 2.31	46.04 ± 4.02	44.00 ± 5.11
1.0	5	**41.49** ± 2.33	**40.04** ± 1.42	**50.33** ± 2.65	**48.91** ± 4.27
2.5	2	39.60 ± 2.30	37.92 ± 2.45	44.93 ± 2.53	44.42 ± 2.70

Effect of Temporal Window Size. Table A.7 reports the effect of temporal window size on both the EEG student and teacher on EAV. With a 0.1s window, the student achieves 38.67% accuracy and the teacher 43.90%, indicating that overly short segments may provide insufficient spectral information for stable EEG representation learning. Increasing the window size to 0.5s improves both models, with the student reaching 40.50% and the teacher 46.04%. The 1.0s window achieves the best overall performance, yielding 41.49% accuracy for the student and 50.33% for the teacher, suggesting a good balance between spectral representation quality and temporal granularity. When the window size increases to 2.5s, performance decreases to 39.60% and 44.93%, respectively. With only two temporal windows per trial, the reduced temporal resolution limits the ability of EDD to distinguish emotionally informative moments. Overall, these results suggest that a 1.0s window provides the most suitable temporal resolution for EXIST on EAV.

Per-class Analysis via Confusion Matrices Fig. A.2 compares the per-class recognition patterns of the EEG student before and after EXIST distillation on one cross-subject split of the EAV dataset. Without distillation, Happiness is the best-recognised class (466/720, 64.7%), while Sadness is the most confused, with 255 samples misclassified as Neutral. After distillation, several consistent improvements emerge. Anger recognition improves from 202 to 233 correct predictions, and Calmness improves substantially from 194 to 301, suggesting that EXIST helps the student develop more discriminative representations for these two categories. Happiness remains the strongest class (478). Sadness remains the most challenging category, with 214 samples still confused as Neutral, consistent with the known difficulty of separating negative low-arousal states in cross-subject EEG. Overall, distillation produces more balanced recognition across categories.

Analysis of the Source of Improvement To analyze whether the improvement of EXIST comes from the regularization effect of the distillation objectives or the privileged audiovisual (AV) knowledge provided by the teacher, we conduct an additional ablation experiment. Specifically, we remove the AV modality from the teacher while keeping the proposed CDD and EDD objectives unchanged, resulting in an EEG-only self-distillation variant. In this setting, the teacher generates supervision using only EEG representations without accessing privileged AV information. As shown in Table A.8, the EEG-only self-distillation variant achieves 38.39% Acc and 38.01% F1 on EAV, which



(a) W/O Distillation (b) W/ Distillation

Figure A.2: Confusion matrices of the EEG student without (a) and with (b) EXIST distillation on EAV. Numbers indicate the count of test samples per cell.

Table A.8: Ablation study on the source of teacher supervision. The EEG-only self-distillation variant removes AV information while keeping the distillation objectives unchanged.

Method	Teacher Input	Acc (%)	F1 (%)
No distillation	–	38.72 $\pm$ 2.58	38.04 $\pm$ 2.55
EEG self-distillation	EEG	38.39 $\pm$ 3.02	38.01 $\pm$ 2.93
EXIST	EEG+AV	41.49 $\pm$ 2.33	40.04 $\pm$ 1.42

is comparable to the student without distillation (38.72% Acc and 38.04% F1), but lower than EXIST (41.49% Acc and 40.04% F1). This result indicates that the distillation objectives alone provide limited improvement when no additional modal information is available. The performance gain of EXIST comes mainly from the privileged AV guidance, which transfers complementary emotional information from the multimodal teacher to the only student with EEG.

Effect of EEG Backbone Scale. We first investigate the effect of EEG backbone capacity on the distilled student performance. As shown in Table A.9, increasing the GCN/LSTM dimensions from 64/128 to 128/256 increases the parameter count from 1.05M to 8.71M, but does not lead to further performance improvement. The smaller backbone achieves 41.49% Acc and 40.04% F1, while the larger backbone obtains 41.25% Acc and 38.43% F1. This result suggests that the performance gain of EXIST does not mainly come from increasing student model capacity, but from the proposed knowledge transfer mechanism. The performance drop of the larger model may result from overfitting due to the limited dataset size.

Analysis of Teacher–Student Knowledge Alignment. To further analyze the knowledge transfer process, we measure the KL divergence between the EEG teacher and student representations before and after distillation on the EAV test set. As shown in Table A.10, applying CDD reduces the cortical-dependency KL divergence from 3.17 to 1.16, while applying EDD reduces the emotional-dynamics KL divergence from 1.18 to 0.14. These results indicate that CDD and EDD effectively reduce the discrepancy between teacher and student with respect to their corresponding spatial and temporal supervision targets. The remaining cortical-dependency KL divergence also indicates that knowledge transfer is effective but incomplete, which may partially explain the performance gap between the AV-guided teacher and the EEG-only student. AV can reveal substantially richer affective structure in EEG space, but directly distilling this privileged knowledge is difficult. EXIST converts this information into EEG-native spatial and temporal supervision.

Cross-backbone Transferability Analysis To investigate whether EXIST is tied to a specific EEG backbone, we further evaluate the proposed CDD and EDD objectives on EEGFormer (Wan et al., 2023), a Transformer-based EEG emotion recognition model with an architecture different from

Table A.9: Effect of EEG backbone parameter scale on distilled student performance.

GCN/LSTM Dims	Params	Acc (%)	F1 (%)
64/128	1.05M	41.49 ± 2.33	40.04 ± 1.42
128/256	8.71M	41.25 ± 1.90	38.43 ± 2.48

Table A.10: Analysis of teacher–student knowledge alignment. CDD and EDD improve the alignment of cortical dependency and emotional dynamics, respectively.

Method	Cortical KL ↓	Emotional KL ↓	Acc (%)	F1 (%)
Student w/o KD	3.17	1.18	38.72	38.04
EXIST	**1.16**	**0.14**	**41.49**	**40.04**

Table A.11: Cross-backbone evaluation on EEGFormer.

Backbone	Distillation	EAV Acc (%)	EAV F1 (%)	PME4 Acc (%)	PME4 F1 (%)
EEGFormer	None	38.08 ± 0.48	35.58 ± 0.82	25.15 ± 0.15	16.90 ± 1.70
EEGFormer	EXIST (CDD+EDD)	39.45 ± 0.41	38.28 ± 0.37	27.68 ± 0.53	26.18 ± 0.20

our default GCN–LSTM student. EEGFormer consists of two key components: a Synchronous Transformer that captures inter-electrode dependencies through spatial attention, and a Temporal Transformer that models temporal interactions among EEG tokens. Therefore, its intermediate representations naturally align with the two types of knowledge transferred by EXIST: CDD targets spatial dependency patterns among EEG electrodes, while EDD targets temporal importance distributions. For this experiment, we replace only the EEG student backbone with EEGFormer while keeping the teacher network, distillation objectives, and optimization protocol unchanged. As shown in Table A.11, EXIST consistently improves EEGFormer on both datasets. These results demonstrate that EXIST is not restricted to the original GCN–LSTM backbone. Instead, CDD and EDD can be instantiated on different EEG architectures as long as the backbone provides spatial and temporal intermediate representations.