CoSim: Unleashing Eye Movements for EEG-Free Emotion Recognition via Conditional Prompting and Similarity-Guided Augmentation

Xiaoling An[1], Siyuan He[1], Zewen Liu[1], Chaofei Zhou[1], Jing Qin[2], Daoqiang Zhang[3], and Xiaoke Hao[1,4(✉)]

[1] School of Artificial Intelligence, Hebei University of Technology, Tianjin, China
[2] School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China
[3] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[4] Haihe Laboratory of Brain-computer Interaction and Human-machine Integration, Tianjin, China
haoxiaoke@hebut.edu.cn

Abstract. Emotion recognition using physiological signals is pivotal for mental health assessment, and multimodal fusion is a key strategy for enhancing performance. Electroencephalography (EEG) and eye movement (EYE) signals are widely used as they provide complementary affective information. However, EEG's complex acquisition limits its real-world feasibility, while the more accessible EYE modality exhibits inferior performance when used alone. To address this, we propose CoSim, a framework integrates Conditional Prompting with Similarity-Guided Knowledge Augmentation (SKA). Our approach aims to enhance the inference-time performance of the EYE modality by transferring knowledge from the information-rich EEG modality, which is available exclusively for training. During training, a shared representation space is learned from paired EEG-EYE data, augmented by samples from similar subjects to mitigate data scarcity and individual variability. During testing, when EEG is unavailable, a Conditional Prompting Generator (CPG) generates a discriminative representation from the EYE input to compensate for the missing modality. The framework is built on a Fourier-Enhanced Transformer (FET) backbone to capture critical time-frequency characteristics. Extensive experiments demonstrate that our model achieves state-of-the-art performance. For instance, on the SEED dataset, our EYE-only model reaches 98.52% accuracy, achieving comparable performance to full-modality and validating its potential for practical deployment. The code is available at https://github.com/AnXiaoL/CoSim.

Keywords: Missing modality · cross-modal emotion recognition · EEG · eye movement · conditional prompting · personalized modeling.

1 Introduction

Multimodal fusion of physiological signals, particularly electroencephalography (EEG) and eye movement (EYE), is a cornerstone for achieving state-of-the-art

emotion recognition [26,16,27,25]. However, a significant gap exists between this laboratory-optimized paradigm and its real-world clinical use. While EEG offers rich neural information, its complex acquisition process limits its feasibility [21], especially for sensitive populations, such as children. In contrast, EYE signals offer high accessibility and potential for large-scale deployment [13,6,5]. This discrepancy raises a critical scientific challenge: How can we effectively reconstruct the discriminative information inherent to the EEG modality using only EYE data, to achieve performance comparable to that of full-modality fusion?

To address this challenge, prior research has primarily followed two paths. The first focuses on feature generation, using methods from early regression models [7] to more recent Conditional Generative Adversarial Networks [23,20]. The other centers on cross-modal representation alignment, which learns a shared latent space using techniques like Deep Canonical Correlation Analysis [2] or contrastive learning [8,12].

Despite progress, critical challenges remain. Existing methods often struggle to generate features that are semantically aligned with the available modality and typically overlook valuable cross-subject information. These limitations severely hinder performance, especially in scenarios with limited data or significant distribution shifts. To address these limitations, we propose CoSim, a framework that integrates Conditional Prompting with Similarity-Guided Augmentation. Our main contributions are as follows:

- Conditional Prompting Generator (CPG): We introduce a novel prompting mechanism that generates pseudo-EEG representations conditioned on the EYE input. This ensures the synthesized features are semantically relevant, moving beyond traditional generation paradigms to reconstruct truly relevant information.
- Similarity-Guided Knowledge Augmentation (SKA): We propose a strategy to augment the training data by leveraging knowledge from physiologically similar subjects. This creates a more robust and personalized training process, directly tackling the issue of limited individual data.
- Fourier-Enhanced Transformer (FET) Backbone: We employ a FET backbone to effectively model critical physiological cues. While standard Transformers [18] lack an inductive bias toward frequency features, we recognize that even without the explicit neural oscillations of EEG, EYE signals contain physiological, quasi-periodic patterns (e.g., saccade intervals) correlated with emotional states [17]. The FET backbone is specifically chosen to help capture these subtle but vital time-frequency characteristics through its structural design, leading to state-of-the-art performance.

2 Method

2.1 Overview

As illustrated in Fig. 1, our framework consists of three stages: pretrain, finetune, and test.

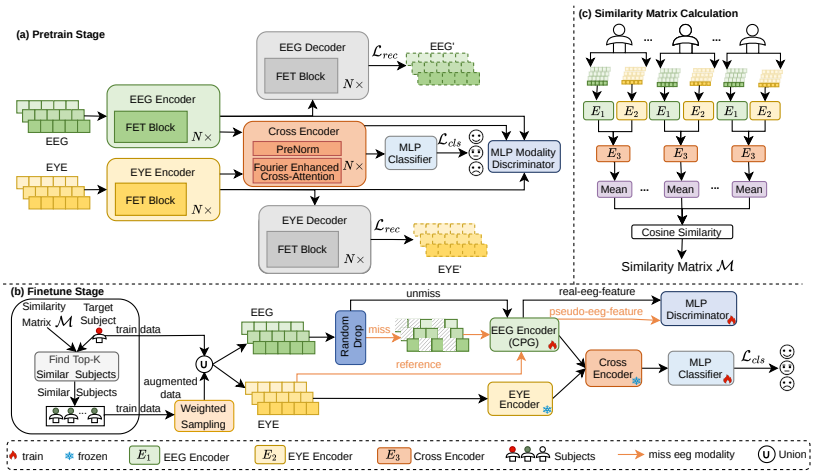

Fig. 1. The overall framework of CoSim.

Universal Multimodal Pretrain. We first pretrain a dual-stream architecture on paired EEG-EYE data from all subjects to learn a robust multimodal representation space. The primary classification task, optimized via Cross-Entropy loss ($\mathcal{L}_{\text{cls}}$), is regularized by two auxiliary objectives: Feature Reconstruction, using an L1 loss ($\mathcal{L}_{\text{rec}}$) to preserve modality-specific details, and Adversarial Modality Alignment ($\mathcal{L}_{\text{madv}}$), which uses a standard adversarial loss to enforce the learning of modality-invariant features. The total pretraining loss is $\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{madv}}\mathcal{L}_{\text{madv}}$, where λ_{rec} and λ_{madv} are trade-off hyperparameters. The trained encoders and the fusion module provide a strong initialization for the next stage.

Personalized Adaptive Finetune. To adapt the model for a target individual, we finetune the CPG and the classifier. First, the target's training data is enriched via SKA. We then employ Random Modality Dropout by randomly masking the EEG input. When EEG is dropped, the CPG generates a pseudo-EEG representation from the EYE input. The model is then optimized with a combined loss $\mathcal{L}_{\text{ft}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}$, where λ_{adv} balances the adversarial loss $\mathcal{L}_{\text{adv}}$, which ensures the generated features distributionally match real EEG features. To ensure model robustness and prevent degraded generalization, most pretrained parameters are frozen.

Test. At the test stage, only the target's EYE signal is available. The finetuned CPG generates a pseudo-EEG representation from the EYE input. Both real EYE and pseudo-EEG representations are then fed into the fusion module for the final emotion classification.

2.2 Conditional Prompting Generator

To explicitly address the absence of EEG during inference, we introduce CPG within the EEG encoder, as illustrated in Fig. 2a. When EEG is missing, our prompt is dynamically adapted to the input, comprising a static base prompt (P_{base}) and a conditional prompt (P_{cond}). P_{base} is a learnable shared vector capturing the general distribution prior of the EEG modality. In contrast, P_{cond} is generated by an MLP-based module conditioned on the current EYE representation X_{EYE}.

$$P_{\text{cond}} = \text{MLP}(\text{MeanPool}(X_{\text{EYE}})). \tag{1}$$

The final prompt P is the sum of these two components:

$$P = P_{\text{base}} + P_{\text{cond}}. \tag{2}$$

This combined prompt is then processed through N FET Blocks to generate the final pseudo-EEG feature Z^p_{EEG}. This design ensures the generated features simultaneously adhere to the general EEG distribution via P_{base} and align with the emotional semantics of the current EYE input via P_{cond}. To ensure the quality of the generated features, we employ feature-level adversarial learning [4]. A discriminator D is trained to distinguish between real EEG features Z^r_{EEG} and the generated pseudo-features Z^p_{EEG}. The discriminator and generator losses are defined as:

$$\mathcal{L}_{\text{dis}} = -\mathbb{E}[\log D(Z^r_{\text{EEG}})] - \mathbb{E}[\log(1 - D(Z^p_{\text{EEG}}))], \tag{3}$$

$$\mathcal{L}_{\text{gen}} = -\mathbb{E}[\log D(Z^p_{\text{EEG}})]. \tag{4}$$

Through this adversarial process, the generator learns an effective data-driven cross-modal mapping from EYE to EEG, producing pseudo-features sufficient for the subsequent fusion and classification tasks.

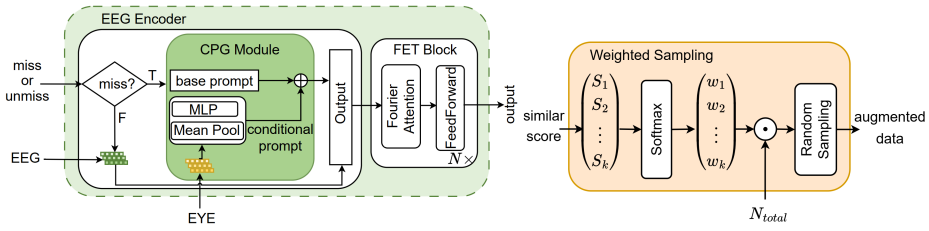

Fig. 2. (a) The overall framework of EEG Encoder (CPG). (b) The detailed process of Weighted Sampling.

2.3 Similarity-Guided Knowledge Augmentation

The CPG often suffers from degraded generalization when a target subject's data is limited or distributionally biased. To mitigate this, we propose the SKA strategy that aims to regularize the training process. Specifically, SKA identifies subjects with comparable emotion-physiology feature distributions and uses their data to augment the target's training set. Unlike conventional neighbor-based methods, SKA utilizes joint EEG-EYE features to ensure deep representation matching rather than mere volume expansion. To achieve this, we first quantify inter-subject similarity using fused multimodal features from a pretrained model. For each subject S_i, we compute a prototype vector $\mathbf{v}_i$ by averaging the features of all training samples for that subject. This vector represents the subject's characteristic emotion-physiology response pattern. The similarity between subjects S_i and S_j is then calculated using cosine similarity:

$$\mathrm{Sim}(S_i, S_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|\|\mathbf{v}_j\|}. \tag{5}$$

This constructs a subject similarity matrix $\mathcal{M}$ where $\mathcal{M}_{i,j} = \mathrm{Sim}(S_i, S_j)$. Based on $\mathcal{M}$, we select the top-K most similar subjects for a target subject S_{target}. We then perform weighted sampling to augment the training data (as shown in Fig. 2b), with weights assigned via a Softmax function to prioritize more similar subjects:

$$w_k = \frac{\exp(\mathrm{Sim}(S_{\text{target}}, S_k)/\tau)}{\sum_{m=1}^{K} \exp(\mathrm{Sim}(S_{\text{target}}, S_m)/\tau)}. \tag{6}$$

Here, τ is a temperature hyperparameter. For each source subject S_k, we randomly draw $N_k = \lfloor N_{\text{total}} \cdot w_k \rfloor$ samples, where N_{total} is the total augmentation size. This weighted sampling approach enriches data diversity while minimizing distribution shift from irrelevant subjects.

2.4 Fourier-Enhanced Transformer Backbone

The representational capacity of the backbone is critical for extracting robust features. To better capture the time-frequency patterns inherent in physiological signals, we employ a FET backbone. Inspired by Fourier Analysis Networks (FAN) [1], FET replaces linear projections with a Fourier-Analysis Layer (FANLayer), which models periodic and aperiodic signal components separately. For an input $\mathbf{x} \in \mathbb{R}^d$, the FANLayer projects it into two complementary subspaces:

$$\mathbf{z}_{\text{per}} = \mathbf{W}_{\text{per}}\mathbf{x} + \mathbf{b}_{\text{per}} \quad \text{(Periodic component)}, \tag{7}$$

$$\mathbf{z}_{\text{aper}} = \phi(\mathbf{W}_{\text{aper}}\mathbf{x} + \mathbf{b}_{\text{aper}}) \quad \text{(Aperiodic component)}. \tag{8}$$

The periodic component is then mapped to a trigonometric space and concatenated with the aperiodic one:

$$\mathrm{FAN}(\mathbf{x}) = \mathrm{Concat}\big(\cos(\mathbf{z}_{\text{per}}),\ \sin(\mathbf{z}_{\text{per}}),\ \mathbf{z}_{\text{aper}}\big). \tag{9}$$

By applying this structure to the Query (Q), Key (K), and Value (V) projections, FET creates a frequency-aware representation space. Consequently, the self-attention mechanism can evaluate similarity based on both temporal content and frequency structure:

$$\mathbf{Q} = \mathrm{FAN}(\mathbf{x}), \quad \mathbf{K} = \mathrm{FAN}(\mathbf{x}), \quad \mathbf{V} = \mathrm{FAN}(\mathbf{x}), \tag{10}$$

$$\mathrm{Attn}(\mathbf{x}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_{\mathrm{h}}}}\right)\mathbf{V}. \tag{11}$$

This structural decoupling of periodic and aperiodic components introduces a crucial frequency-domain inductive bias that naturally aligns with the intrinsic spectral properties of physiological signals. By disentangling these components, the model can effectively capture emotion-relevant patterns, thereby improving feature robustness and overall performance.

3 Experiments

3.1 Experimental Settings

Datasets. We validate on two public benchmarks: SEED [28] and SEED-IV [27]. SEED contains three-class emotion data from 12 subjects across three sessions (15 videos per session), while SEED-IV includes four-class data from 15 subjects across three sessions (24 videos per session). Both provide synchronized EEG (62 channels) and EYE signals. Following standard protocols, we adopt a subject-dependent evaluation with train/test video splits of 9:6 for SEED and 16:8 for SEED-IV. For feature extraction, we use 310-dimensional (62 channels $\times$ 5 bands) Differential Entropy (DE) for EEG, and statistical EYE features (33 dimensions for SEED, 31 for SEED-IV).

Implementation Details. To ensure fair comparison, we align our experimental settings (e.g., train/test splits) with prior studies [27,23,20]. The FET backbone has 3 layers, 3 attention heads, a hidden dimension of 96, and a temporal window size of 5. The Random Modality Dropout rate is 0.6. We employ the Adam optimizer, with a learning rate searched in $[10^{-4}, 10^{-1}]$ and a weight decay in $[10^{-4}, 10^{-2}]$. Batch sizes are set to 128 for pretraining and 64 for finetuning. All experiments are implemented in PyTorch 2.0 on two A30 GPUs.

3.2 Performance Comparison

Comparison Methods. We benchmark CoSim against three categories of baselines, all evaluated under the same subject-dependent protocol for a fair comparison. (1) Cross-modal methods [23,20,8,15,10], which share our train-multimodal, test-unimodal setting. (2) Full-modality methods [27,25,11,22,24,14], which use all data and provide a performance upper bound. (3) Unimodal baselines: including an EYE-only model [24] as a performance lower bound and strong EEG-only

Table 1. Cross-Modal Baselines on the SEED and SEED-IV datasets. Results are reported as mean accuracy ± standard deviation (%).

Method	SEED	SEED-IV
Yan et al. [23]	81.02 ± 8.04	75.74 ± **6.66**
ECO-FET [8]	93.69 ± 8.22	87.76 ± 9.19
CLIP [15]	91.67 ± 10.15	86.74 ± 10.74
CANDA [20]	96.60 ± 6.17	90.48 ± 7.33
UMAP [10]	89.32 ± 12.44	80.75 ± 14.21
CoSim	**98.52** ± **4.56**	**94.99** ± 6.77

Table 2. Full-Modality Baselines on the SEED and SEED-IV datasets. Results are reported as mean accuracy ± standard deviation (%).

Method	SEED	SEED-IV
BDAE [27,25]	91.00 ± 8.9	85.10 ± 11.8
DCCA [11]	94.60 ± 6.2	87.50 ± 9.2
DCCA-FCP [22]	95.08 ± 6.42	- -
DTCA [24]	**99.15** ± **2.9**	**99.64** ± **0.5**
MambaMER [14]	96.82 ± 5.2	94.93 ± 6.12
CoSim	98.52 ± 4.56	94.99 ± 6.77

baselines [24,3,9,29,19]. The latter is crucial for assessing whether CoSim, using solely EYE inputs, can effectively compensate for missing information and approach or surpass the performance of strong EEG-only models.

Table 3. Unimodal Baselines on the SEED and SEED-IV datasets. Results are reported as mean accuracy ± standard deviation (%).

Method	SEED	SEED-IV
TAHAG(EEG) [3]	93.11 ± 8.17	89.50 ± 11.75
PGCN(EEG) [9]	96.93 ± 5.11	82.24 ± 14.85
PR-PL(EEG) [29]	94.84 ± 9.16	83.33 ± 10.61
FAT(EEG) [19]	92.10 ± 6.46	82.30 ± 13.21
DTCA(EEG) [24]	96.68 ± 6.5	94.19 ± **3.8**
DTCA(EYE) [24]	84.33 ± 12.4	89.71 ± 7.3
CoSim	**98.52** ± **4.56**	**94.99** ± 6.77

Table 4. Ablation Study on the SEED and SEED-IV datasets. Results are reported as mean accuracy ± standard deviation (%).

Method	SEED	SEED-IV
w/o FET	97.30 ± 7.1	94.64 ± 7.93
w/o SKA	96.45 ± 7.31	93.58 ± 7.77
w/o CPG	84.90 ± 11.41	92.82 ± **6.74**
CoSim	**98.52** ± **4.56**	**94.99** ± 6.77

Main Results. As shown in Table 1, Table 2 and Table 3, CoSim demonstrates clear superiority. It significantly outperforms all cross-modal baselines that operate under the same test-unimodal setting. More critically, our EYE-only model not only closes the gap with strong EEG-only models but surpasses them. For instance, on SEED, CoSim achieves 98.52% accuracy, exceeding the state-of-the-art DTCA (EEG) baseline (96.68%) and nearly matching the full-modality DTCA (99.15%). This validates our framework's ability to successfully compensate for missing EEG information using only EYE signals during inference.

3.3 Ablation Studies and Analysis

Ablation Study. Table 4 evaluates the proposed modules. The results show that removing the CPG caused the largest performance drop (98.52% to 84.90%), as the model lost its ability to reconstruct pseudo-EEG features and had to rely

solely on less discriminative EYE features. Similarly, disabling SKA also led to a consistent accuracy decrease, indicating that the model suffers from degraded generalization without diverse data from similar subjects. Furthermore, replacing FET with a standard Transformer degraded performance by losing the crucial frequency-domain bias needed to capture subtle periodic EYE patterns.

Beyond delivering a significant accuracy boost, our framework maintains practical computational efficiency for edge deployment. On the SEED dataset, the CPG introduces merely 2.14 ms of additional latency (3.22 ms total per sample) with a parameter count of 1.5M. This acceptable overhead is fully justified by the substantial accuracy gains over the baseline.

Generated Feature Visualization Analysis. To validate our CPG module, we used t-SNE to project the real and generated pseudo-EEG features into a 2D space (Fig. 3a and Fig. 3b). The significant overlap between the two distributions indicates a successful alignment on the natural feature manifold of real EEG data. This distributional alignment, combined with the high classification accuracy (Table 1) and the ablation result (Table 4), confirms that our CPG effectively reconstructs the necessary discriminative information from the missing modality, avoiding mode collapse into a static mean pattern.

Parameter Sensitivity Analysis. Our hyperparameter analysis for the SKA module on the SEED dataset (Fig. 3c) reveals a "rise-then-fall" pattern for both the number of similar subjects K and total augmentation size N_{total}. Performance initially improves as incorporating diverse inter-subject knowledge enriches the feature space. However, excessively large values lead to negative transfer from dissimilar subjects or overwhelm the target's original distribution. This highlights a critical trade-off between knowledge diversity and domain shift. For the SEED dataset (467 samples/subject), the optimal balance was empirically found at $K = 2$ and $N_{\text{total}} = 467$.

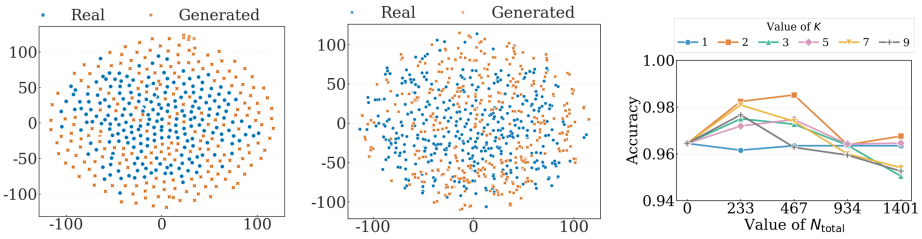

Fig. 3. (a) Feature Distribution on SEED. (b) Feature Distribution on SEED-IV. (c) Analysis of SKA Hyperparameters: K and N_{total}.

4 Conclusion

In this work, we proposed CoSim to enable practical, EEG-free emotion recognition. It bridges the gap with full-modality systems by integrating a Conditional Prompting Generator to reconstruct missing EEG features from EYE signals, Similarity-Guided Knowledge Augmentation to enrich limited data using physiologically similar subjects, and a Fourier-Enhanced Transformer to capture critical time-frequency patterns. On the SEED and SEED-IV datasets, CoSim sets a new state-of-the-art for the EYE-only inference paradigm, paving the way for accessible and low-cost mental health monitoring tools. Future work will explore dynamic EYE features (e.g., scanpaths) and large-scale multi-center datasets to further verify broad-scenario generalization, ultimately moving towards a robust subject-independent framework to maximize clinical utility.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (No. 62576133), in part by the Natural Science Foundation of Hebei Province (No. H2023202901), in part by the Beijing-Tianjin-Hebei Basic Research Cooperation Project (No. J230040).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Dong, Y., Li, G., Tao, Y., Jiang, X., Zhang, K., Li, J., Deng, J., Su, J., Zhang, J., Xu, J.: FAN: Fourier analysis networks. arXiv preprint arXiv:2410.02675 (2024)
2. Fei, C., Li, R., Zhao, L.M., Li, Z., Lu, B.L.: A cross-modality deep learning method for measuring decision confidence from eye movement signals. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 3342–3345. IEEE (2022)
3. Gong, P., Zhou, Y., Huang, S., Wang, P., Zhang, D.: TAHAG: Two-stage domain adaptation with hybrid adaptive graph learning for EEG emotion recognition. IEEE Transactions on Affective Computing **16**(4), 2748–2761 (2025)
4. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. vol. 27 (2014)
5. Grüner, M., Ansorge, U.: Mobile eye tracking during real-world night driving: A selective review of findings and recommendations for future research. Journal of Eye Movement Research **10**(2), 8 (2017)
6. Hwang, A.D., Wang, H.C., Pomplun, M.: Semantic guidance of eye movements in real-world scenes. Vision Research **51**(10), 1192–1205 (2011)
7. Jiang, H., Guan, X., Zhao, W.Y., Zhao, L.M., Lu, B.L.: Generating multimodal features for emotion classification from eye movement signals. Aust. J. Intell. Inf. Process. Syst. **15**(3), 59–66 (2019)
8. Jiang, W.B., Li, Z., Zheng, W.L., Lu, B.L.: Functional emotion transformer for EEG-assisted cross-modal emotion recognition. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1841–1845. IEEE (2024)
9. Jin, M., Du, C., He, H., Cai, T., Li, J.: PGCN: Pyramidal graph convolutional network for EEG emotion recognition. IEEE Transactions on Multimedia **26**, 9070–9082 (2024)

10. Li, Z., Zheng, W.L., Lu, B.L.: Multimodal emotion recognition with missing modality via a unified multi-task pre-training framework. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 5717–5725 (2025)
11. Liu, W., Qiu, J.L., Zheng, W.L., Lu, B.L.: Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Transactions on Cognitive and Developmental Systems **14**(2), 715–729 (2021)
12. Liu, Y.K., Cai, J., Lu, B.L., Zheng, W.L.: Multi-to-single: Reducing multimodal dependency in emotion recognition through contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1438–1446 (2025)
13. Merzon, L., Pettersson, K., Aronen, E.T., Huhdanpää, H., Seesjärvi, E., Henriksson, L., MacInnes, W.J., Mannerkoski, M., Macaluso, E., Salmi, J.: Eye movement behavior in a real-world virtual reality task reveals ADHD in children. Scientific Reports **12**(1), 20308 (2022)
14. Ping, X., Huang, W., Zheng, Y.: MambaMER: Adaptive EEG-guided multimodal emotion recognition with Mamba. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 333–343. Springer (2025)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
16. Soleymani, M., Pantic, M., Pun, T.: Multimodal emotion recognition in response to videos. IEEE Transactions on Affective Computing **3**(2), 211–223 (2011)
17. Staudigl, T., Hartl, E., Noachtar, S., Doeller, C.F., Jensen, O.: Saccades are phase-locked to alpha oscillations in the occipital and medial temporal lobe during successful memory encoding. PLoS Biology **15**(12), e2003404 (2017)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
19. Wang, J., Huang, Y., Song, S., Wang, B., Su, J., Ding, J.: A novel Fourier adjacency transformer for advanced EEG emotion recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2025)
20. Wang, Y., Liu, J.W., Lu, B.L., Zheng, W.L.: From EEG to eye movements: Cross-modal emotion recognition using constrained adversarial network with dual attention. IEEE Transactions on Affective Computing **16**(3), 1543–1556 (2024)
21. Wu, R., Wang, H., Chen, H.T.: A comprehensive survey on deep multimodal learning with missing modality. arXiv preprint arXiv:2409.07825 (2024)
22. Wu, X., Zheng, W.L., Li, Z., Lu, B.L.: Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. Journal of Neural Engineering **19**(1), 016012 (2022)
23. Yan, X., Zhao, L.M., Lu, B.L.: Simplifying multimodal emotion recognition with single eye movement modality. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1057–1063 (2021)
24. Zhang, X., Shi, E., Yu, S., Zhang, S.: DTCA: dual-branch transformer with cross-attention for EEG and eye movement data fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 141–151. Springer (2024)
25. Zhao, L.M., Li, R., Zheng, W.L., Lu, B.L.: Classification of five emotions from EEG and eye movement signals: complementary representation properties. In: 2019 9th

International IEEE/EMBS Conference on Neural Engineering (NER). pp. 611–614. IEEE (2019)
26. Zheng, W.L., Dong, B.N., Lu, B.L.: Multimodal emotion recognition using EEG and eye tracking data. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5040–5043. IEEE (2014)
27. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: Emotionmeter: A multimodal framework for recognizing human emotions. IEEE Transactions on Cybernetics **49**(3), 1110–1122 (2018)
28. Zheng, W.L., Lu, B.L.: 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)
29. Zhou, R., Zhang, Z., Fu, H., Zhang, L., Li, L., Huang, G., Li, F., Yang, X., Dong, Y., Zhang, Y.T., et al.: PR-PL: A novel prototypical representation based pairwise learning framework for emotion recognition using EEG signals. IEEE Transactions on Affective Computing **15**(2), 657–670 (2023)