Trial-Consistent Learning Improves Reliability in EEG Emotion Classification

Leyu Zhou[1,*], **Hongjun Liu**[1,*], **Kuanjian Tang**[1], **Jingyi Hou**[1], **Chao Yao**[1,†]
u202343397@xs.ustb.edu.cn, ustb_lhj@163.com, u202342550@xs.ustb.edu.cn, houjingyi@ustb.edu.cn,
yaochao@ustb.edu.cn
[1]University of Science and Technology Beijing, Beijing, China
[*]Equal contribution. [†]Corresponding author.

Abstract

In stimulus-driven EEG emotion classification, labels are defined at the trial level. Each trial corresponds to a fixed stimulus, and windowing yields multiple dependent repeated measurements within the same trial. However, standard pipelines treat windows as independent samples, which ignores the trial as the true unit and encourages models to latch onto transient noise, leading to inconsistent within-trial predictions and unstable generalization. We turn the trial-as-unit prior in stimulus-driven emotion EEG into a trial-defined structural constraint. Specifically, windows are grouped by their trial IDs and a loss is added that makes predictions for windows within the same trial agree with each other. This penalizes reliance on window-specific transient noise and encourages the model to encode stable, trial-level representations. Across SEED and SEED-IV and three backbones including EEGNet, EEGConformer, and CTNet, the proposed constraint increases within-trial prediction consistency and stabilizes validation behavior. It also improves cross-session classification accuracy.

Keywords: EEG emotion classification; repeated measurements; reliability; trial consistency

Introduction

EEG-based emotion recognition has been widely studied and integrated into affective computing systems for emotion-aware human–computer interaction, such as adaptive multimedia recommendation and content evaluation, user-state monitoring for learning and workload management, and affect assessment in mental-health–related applications (Abdul et al., 2018; Picard, 1997; Schlicher et al., 2025). The standard stimulus-driven pipeline performs window-level classification on short EEG segments, for example 4-second windows extracted in real time from continuous recordings (Li et al., 2023; Zheng & Lu, 2015). Nevertheless, EEG recordings remain highly noise-prone and non-stationary due to artifacts and physiological fluctuations, so practical emotion recognition requires stable representations and reliable predictions, especially when generalizing across sessions and subjects.

Prior works on improving the stability of EEG emotion recognition typically introduces auxiliary losses during training, which can be grouped into three categories. First, consistency and smoothing approaches enforce invariance of predictions or representations under perturbations and augmentations (Laine & Aila, 2017; Tarvainen & Valpola, 2017; Wang et al., 2018). Second, contrastive and self-supervised representation learning methods learn robust embeddings that capture stable temporal structure and reduce reliance on noisy supervision (Alghamdi et al., 2025; Chen et al., 2020). Third, domain generalization and alignment methods address session-to-session and subject-to-subject shifts by encouraging domain-invariant representations through distribution matching or adversarial objectives (Arjovsky et al., 2019; Ganin & Lempitsky, 2015; Sun & Saenko, 2016). Despite the progress of augmentation consistency and cross-domain alignment, we further ask (1) whether stability objectives can explicitly suppress transient, window-specific noise so that predictions remain coherent within a trial, and (2) whether the trial-defined repeated-measurement structure can be leveraged as a training signal to improve cross-session robustness without introducing any test-time dependence on metadata.

Accordingly, we propose a consistency learning principle for stimulus-driven EEG emotion recognition that exploits the trial-as-unit prior while preserving the standard window-level deployment interface. It consists of two key components:

(1) **Trial-Defined Weak Grouping**. In stimulus-driven protocols, each trial corresponds to one stimulus presentation and yields multiple window observations that share the same context and annotation. We use training-set trial identifiers to form weak groups, which expose the repeated-measurement structure without introducing any architectural changes.

(2) **Intra-Trial Predictive Alignment**. Building on this grouping, we augment the standard window-level objective with an intra-trial regularizer that aligns the predictive distributions of windows from the same trial toward a shared trial-level prototype. This discourages the model from fitting window-specific transient cues and promotes trial-consistent representations.

Experiments on SEED and SEED-IV demonstrate improved reliability from three complementary perspectives, including higher intra-trial prediction consistency, more stable training dynamics as measured by reduced epoch-to-epoch variation in validation accuracy, and stronger robustness to task-irrelevant perturbations under test-time Gaussian noise. Overall, exploiting weak trial-level structure via training-only consistency learning yields consistent performance gains and more reliable emotion recognition across datasets and backbones.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Related Work

EEG Emotion Recognition Benchmarks and Architectures

EEG-based emotion recognition is a widely used testbed for affective inference under low signal-to-noise ratio, pronounced non-stationarity, and repeated-measurement variability. This work focuses on SEED and SEED-IV, two stimulus-elicited benchmarks with trial-based organization and repeated recordings (Zheng & Lu, 2015; Zheng et al., 2019). To ensure representative architectural coverage, experiments are conducted with commonly adopted EEG backbones spanning compact CNNs and CNN–Transformer hybrids, including EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2023), and CTNet (Zhao et al., 2024). Rather than introducing a new architecture, the emphasis is on whether a training-only principle can improve reliability in a backbone-agnostic manner.

Robust Learning Under Structured Shift

A substantial literature in EEG affective computing addresses distribution shift across subjects and sessions, often through transfer learning or domain adaptation (Chai et al., 2016). More broadly, robust learning frameworks such as group distributionally robust optimization and invariance-based learning aim to mitigate failures under group-wise shifts, where spurious correlations vary across environments (Arjovsky et al., 2019; Sagawa et al., 2020). Related discussions of "shortcut learning" highlight that deep models may exploit predictive but non-causal cues, leading to inflated in-distribution performance yet poor robustness (Geirhos et al., 2020). In contrast to shifts defined by subjects or sessions, this work targets the trial-defined repeated-measurement structure in stimulus-elicited paradigms, where multiple windows constitute correlated observations of a trial-level latent affective state.

Consistency Regularization

Consistency regularization promotes prediction agreement under perturbations and has been effective for stabilizing learning, especially in semi-supervised settings. Prior methods typically define perturbations through data augmentation, teacher–student targets (e.g., temporal ensembling), or adversarially chosen noise (Miyato et al., 2019; Tarvainen & Valpola, 2017). The key distinction here is that perturbations arise naturally from the data collection process: repeated measurements within a stimulus trial induce structured variability across windows that share the same trial label. Accordingly, trial-defined groups are used as a training-only supervision signal to encourage within-trial predictive agreement and respect within-trial latent-state stability, reducing intra-trial inconsistency and validation oscillation while keeping inference unchanged and free of trial metadata.

Method

Problem Setup, Notation

Stimulus-elicited EEG emotion classification is studied in a *within-subject* setting with a *cross-session* evaluation protocol. In paradigms such as SEED and SEED-IV, recordings are naturally organized into *trials*: each trial corresponds to a continuous stimulus segment (e.g., a video clip) and is assigned a single emotion label. From the standpoint of experimental design, the latent affective state within a trial is expected to be relatively stable at the label level, whereas windowed EEG segments can be viewed as repeated and noisy observations of that trial-level state.

After windowing, the learning problem is formulated at the window level while respecting the session-based split. The training set is

$$\mathcal{D}_{\text{tr}} = \{(\mathbf{x}_i, y_i, g_i)\}_{i=1}^{N}, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^{C\times T}$ denotes the i-th EEG window (in our experiments $C = 62$ channels and $T = 800$ samples), $y_i \in \{1, \ldots, K\}$ is the emotion label, and $g_i \in \{1, \ldots, G\}$ is the trial identifier (trial ID) indicating which trial the window comes from. Crucially, g_i is used *only during training* to construct a regularization term; it is not provided as an input feature to the model, and it is not used during validation or testing.

Taking SEED for example, the raw recordings can be summarized as subject × session × trial × channel × time: each subject completes 3 sessions; each session contains 15 trials; and each trial corresponds to one stimulus segment with a single label. Segmenting the continuous EEG within each trial into fixed-length windows yields samples $\mathbf{x}_i \in \mathbb{R}^{62\times 800}$. This windowing step exposes an additional piece of structure: windows originating from the same trial share the same stimulus context and trial-level label, and thus can be viewed as repeated measurements of a common latent affective state. We refer to this as *trial-defined weak grouping*, and use it only to define training-time regularization.

Trial-Defined Weak Grouping.

Windows sharing the same trial ID form a group

$$\mathcal{I}_g = \{i \mid g_i = g\}, \qquad n_g = |\mathcal{I}_g|. \tag{2}$$

This grouping provides a lightweight, metadata-derived notion of "same-state" windows within each mini-batch, which we exploit to encourage stable within-trial predictive distributions during training.

Backbone Classifier and Supervised Objective

Let $f_\theta(\cdot)$ denote an EEG classifier backbone (e.g., EEGNet, EEGConformer, or CTNet) parameterized by θ. Given an input window $\mathbf{x}_i$, the model outputs logits

$$\mathbf{z}_i = f_\theta(\mathbf{x}_i) \in \mathbb{R}^K, \tag{3}$$

and a predictive distribution via softmax

$$\mathbf{p}_i = \text{softmax}(\mathbf{z}_i), \quad p_{i,k} = \frac{\exp(z_{i,k})}{\sum_{j=1}^{K}\exp(z_{i,j})}. \tag{4}$$

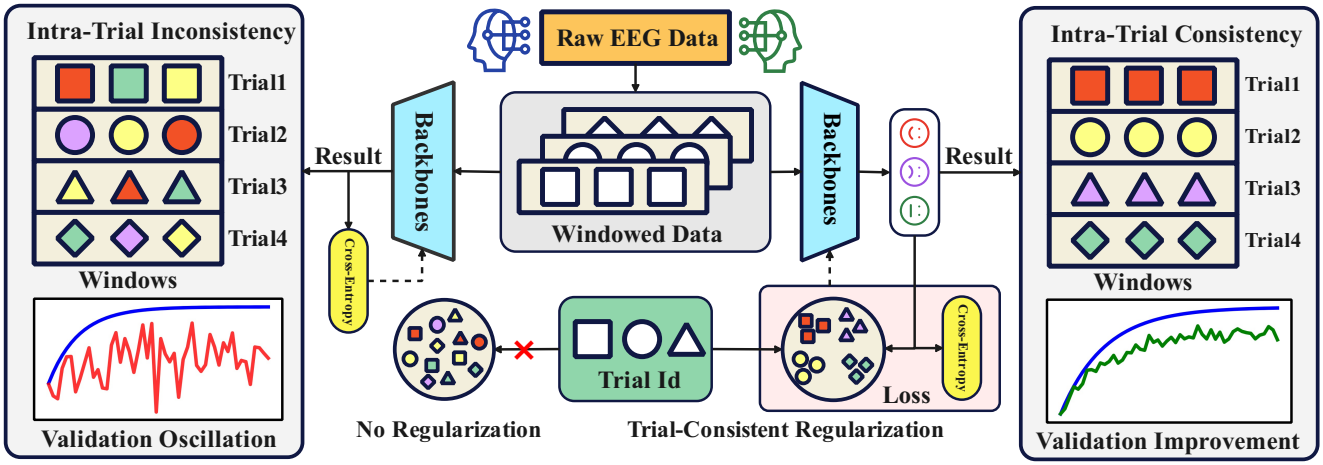

Figure 1: **Conceptual illustration of TrialCons.** In this illustration, colors represent classification results (white indicates unclassified windows), while shapes (e.g., squares, circles) correspond to windows from different trials. The dashed lines depict the loss constraints applied to the backbone model. Left: Traditional window-based training leads to intra-trial inconsistency and validation oscillation. Right: TrialCons aligns predictions within trials, ensuring consistency and stable improvement.

The standard supervised loss is cross-entropy:

$$\mathcal{L}_{\text{cls}}(\theta) = \frac{1}{B}\sum_{i=1}^{B} \ell_{\text{CE}}(\mathbf{z}_i, y_i), \quad \ell_{\text{CE}}(\mathbf{z}_i, y_i) = -\log p_{i,y_i}, \quad (5)$$

where B is the mini-batch size.

Intra-Trial Predictive Alignment

Motivation. Under the trial-stability assumption, windows from the same trial are repeated observations of a shared trial-level latent affective state and therefore should yield similar predictive distributions. Standard window-level ERM ignores this grouping and may overfit transient, task-irrelevant fluctuations. A training-only regularizer is introduced to promote within-trial predictive agreement (Figure 1). To avoid shortcut learning from explicit trial metadata, trial identifiers are never fed to the model as input features; they are used only to define groups for a training-only consistency regularizer.

Trial mean distribution. Consider a mini-batch containing trial IDs. Let $\mathbf{p}_i = \text{softmax}(f_\theta(\mathbf{x}_i)) \in [0,1]^K$ denote the window-level predicted class-probability vector, where $f_\theta(\cdot)$ is the backbone and K is the number of classes. For any trial g appearing in the batch, define its index set and size as

$$\mathcal{I}_g = \{i \in \{1,\dots,B\} \mid g_i = g\}, \qquad n_g = |\mathcal{I}_g|. \quad (6)$$

The mean predictive distribution for trial g in the batch is

$$\bar{\mathbf{p}}_g = \frac{1}{n_g}\sum_{i\in\mathcal{I}_g} \mathbf{p}_i. \quad (7)$$

In implementation, we stop gradients through the batch mean and treat $\bar{\mathbf{p}}_g$ as a target distribution, i.e., we use $\text{stopgrad}(\bar{\mathbf{p}}_g)$ in the consistency term to avoid self-influence through the mean.

KL consistency loss. Disagreement between each window prediction and the (stopped-gradient) trial mean is measured using Kullback–Leibler divergence:

$$D_{\text{KL}}(\mathbf{p}_i \| \text{stopgrad}(\bar{\mathbf{p}}_g)) = \sum_{k=1}^{K} p_{i,k} \log \frac{p_{i,k}}{\text{stopgrad}(\bar{p}_{g,k})}. \quad (8)$$

The batch-level trial-consistency loss averages this divergence over all trials with at least two windows in the batch:

$$\mathcal{L}_{\text{con}}(\theta) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \frac{1}{n_g}\sum_{i\in\mathcal{I}_g} D_{\text{KL}}(\mathbf{p}_i \| \text{stopgrad}(\bar{\mathbf{p}}_g)), \quad (9)$$

where $\mathcal{G} = \{g \mid n_g \geq 2\}$ denotes the set of eligible trials in the batch. Trials appearing only once ($n_g = 1$) are excluded because within-trial agreement cannot be computed. In practice, the effectiveness of $\mathcal{L}_{\text{con}}$ depends on how often multiple windows from the same trial co-occur within a mini-batch.

Batch eligibility for TrialCons. The consistency term $\mathcal{L}_{\text{con}}$ is defined only when a mini-batch contains at least one *eligible* trial, namely a trial with at least two windows in the batch ($n_g \geq 2$). Under default random mini-batch sampling, the number of eligible trials per batch is 5.48 on average (median 6.00) for SEED and 8.28 on average (median 8.00) for SEED-IV, with an activation rate of 1.00 on both datasets. This indicates that $\mathcal{L}_{\text{con}}$ is active in essentially every mini-batch and provides a persistent training signal.

Overall objective

The final training objective is

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{cls}}(\theta) + \lambda\,\mathcal{L}_{\text{con}}(\theta), \quad (10)$$

where $\mathcal{L}_{\text{cls}}$ is the standard cross-entropy classification loss, $\mathcal{L}_{\text{con}}$ is the proposed trial-consistency regularizer, and $\lambda \geq 0$ controls the regularization strength.

Experiments

Experimental Setup

Datasets. Experiments are conducted on two emotion recognition benchmarks: SEED (3-class) and SEED-IV (4-class) (Zheng et al., 2019). Each session contains a fixed set of stimulus trials: SEED has 15 trials per session, while SEED-IV has 24 trials per session.

Windowing. Two segmentation protocols are considered for each dataset: (i) **non-overlap** (stride $s = w$) and (ii) **50% overlap** (stride $s = 0.5w$), where the window length w is fixed to 800 samples. Each window is represented as $\mathbf{x} \in \mathbb{R}^{62\times800}$ (62 EEG channels for both SEED and SEED-IV). After windowing, samples are represented as tensors of shape $\mathbb{R}^{N\times62\times800}$. Under the cross-session within-subject protocol, the resulting numbers of windows are summarized in Table 1.

Table 1: Number of windows (N) in two datasets.

Dataset	Non-overlap		50% overlap	
	Train	Test	Train	Test
SEED	25,260	12,630	50,310	25,155
SEED-IV	25,245	12,330	50,100	24,465

Evaluation protocol. A cross-session within-subject protocol is adopted. Each subject is treated as an independent evaluation unit, where Sessions 1–2 are used for training and Session 3 is used for testing. From Sessions 1–2, 15% of windows are held out for validation and the remaining 85% are used for training. TrialCons is implemented as an additional loss term on top of standard EEG classification backbones, including EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2023), and CTNet (Zhao et al., 2024), with architectures kept unchanged. For each backbone, **ERM** is compared with **ERM + TrialCons**. Trial identifiers are used only within the training split to construct the regularizer, and no trial metadata is used during validation or testing. The checkpoint with the highest validation accuracy is selected for reporting.

Optimization details. All methods share the same optimization protocol within each dataset and backbone setting. To ensure that validation oscillation is not an artifact of an overly aggressive learning rate, we perform a learning-rate sweep for each method and report results under its best validation performance. Unless otherwise specified, we use Adam with a base learning rate of 0.001, batch size 256, and train for 50 epochs; the consistency weight is set to $\lambda = 0.5$. All results are averaged over 5 random seeds.

Metrics

Both predictive performance and reliability are evaluated. Performance is measured by test accuracy. Reliability is characterized by (i) trial consistency (TC), which measures agreement among window-level predictions within the same trial, and (ii) validation total variation (TV), which measures epoch-to-epoch variability of validation accuracy.

Trial consistency. For a trial g with windows $\{x_i\}_{i\in\mathcal{I}_g}$, the predicted label for window i is $\hat{y}_i = \arg\max_c p_\theta(c \mid x_i)$. Trial consistency is defined as the fraction of windows whose predicted labels match the trial's majority prediction,

$$\mathrm{TC}(g) = \frac{1}{|\mathcal{I}_g|}\sum_{i\in\mathcal{I}_g}\mathbb{I}\Big[\hat{y}_i = \mathrm{mode}\Big(\{\hat{y}_j\}_{j\in\mathcal{I}_g}\Big)\Big], \tag{11}$$

and the overall TC is averaged across trials in the evaluation split,

$$\mathrm{TC} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\mathrm{TC}(g). \tag{12}$$

Lower TC indicates larger within-trial disagreement, meaning that windows from the same trial lead to less coherent decisions.

Validation total variation. Let $a_t \in [0,1]$ denote the validation accuracy at epoch $t \in \{1,\dots,T\}$. Validation total variation is defined as

$$\mathrm{TV} = \frac{1}{T-1}\sum_{t=2}^{T}|a_t - a_{t-1}|\,. \tag{13}$$

Larger TV indicates greater epoch-to-epoch variability, while smaller TV indicates a smoother validation trajectory.

Main Results

Table 2 reports cross-session emotion classification results across two datasets, two windowing protocols, and three backbones. Adding TrialCons consistently improves both performance and reliability. Specifically, classification accuracy increases by 3.8–9.4 percentage points across all settings, with the largest gain achieved by CTNet on SEED with 50% overlap (57.3 → 66.7). Trial-level reliability also improves, as TC increases by 0.07–0.20, with the largest improvement on SEED-IV (non-overlap) using CTNet (0.40 → 0.60), indicating substantially more coherent predictions among repeated windows from the same trial. Moreover, validation dynamics become markedly smoother, as TV achieves up to 38% relative reduction, e.g., EEGConformer on SEED-IV with 50% overlap (0.108 → 0.067), suggesting that the optimization process is less sensitive to epoch-to-epoch fluctuations. Overall, the joint gains in accuracy, higher TC, and lower TV support that enforcing trial-consistent predictions reduces reliance on transient window-specific cues and improves generalization under session shifts.

Ablation and Analysis

Hyperparameter analysis. To assess robustness to the consistency weight and justify the default choice, λ is swept in $\{0, 0.1, 0.3, 0.5, 1.0\}$ with all other settings fixed (Table 3).

Table 2: Results under the cross-session within-subject protocol. Best in each column is **bold**; second-best is underlined. Acc/TC: higher is better; TV: lower is better.

Backbone	Method	SEED						SEED-IV					
		Non-overlap			50% overlap			Non-overlap			50% overlap		
		Acc (%)↑	TC↑	TV↓	Acc (%)↑	TC↑	TV↓	Acc (%)↑	TC↑	TV↓	Acc (%)↑	TC↑	TV↓
EEGNet	ERM	53.7 (±0.71)	0.51 (±0.03)	0.183 (±0.022)	55.2 (±0.66)	0.54 (±0.02)	0.227 (±0.019)	33.8 (±0.67)	0.40 (±0.02)	0.095 (±0.019)	34.9 (±0.62)	0.40 (±0.02)	0.068 (±0.018)
	+TrialCons	60.3 (±0.69)	0.63 (±0.01)	0.150 (±0.020)	62.7 (±0.51)	0.65 (±0.02)	0.168 (±0.022)	39.9 (±0.55)	0.48 (±0.02)	0.070 (±0.018)	**40.3** (±0.64)	**0.54** (±0.01)	**0.047** (±0.017)
EEGConf	ERM	54.5 (±0.71)	0.52 (±0.01)	0.165 (±0.025)	53.9 (±0.80)	0.52 (±0.03)	0.115 (±0.020)	33.1 (±0.63)	0.38 (±0.02)	0.103 (±0.019)	33.6 (±0.72)	0.39 (±0.02)	0.108 (±0.020)
	+TrialCons	60.9 (±0.55)	0.60 (±0.02)	**0.149** (±0.019)	61.8 (±0.70)	0.69 (±0.02)	**0.097** (±0.019)	36.9 (±0.62)	0.48 (±0.01)	0.088 (±0.021)	37.8 (±0.63)	0.39 (±0.02)	0.067 (±0.019)
CTNet	ERM	55.8 (±0.70)	0.57 (±0.02)	0.241 (±0.026)	57.3 (±0.65)	0.63 (±0.01)	0.182 (±0.023)	34.1 (±0.64)	0.40 (±0.02)	0.081 (±0.018)	33.9 (±0.59)	0.38 (±0.02)	0.089 (±0.019)
	+TrialCons	**63.6** (±0.63)	**0.68** (±0.01)	0.173 (±0.022)	**66.7** (±0.55)	**0.70** (±0.02)	0.126 (±0.020)	**40.5** (±0.60)	**0.60** (±0.01)	**0.064** (±0.017)	39.1 (±0.48)	0.53 (±0.02)	0.082 (±0.018)

Table 3: Accuracy (%) versus consistency weight λ under the cross-session within-subject protocol using EEGNet. Values are mean ± std across random seeds. Best in each column is **bold**; second-best is underlined.

λ	SEED		SEED-IV	
	Non-overlap	50% overlap	Non-overlap	50% overlap
0.0	53.7±0.7	55.2±0.7	33.8±0.7	34.9±0.6
0.1	53.9±0.6	57.2±0.7	34.5±0.5	36.8±0.6
0.3	58.1±0.6	62.0±0.6	37.6±0.7	38.0±0.6
0.5	**60.3±0.7**	**62.7±0.5**	39.9±0.6	**40.3±0.6**
1.0	59.3±0.6	61.3±0.6	**40.0±0.6**	39.5±0.6

Introducing a small-to-moderate λ consistently improves accuracy over ERM, while overly large values can saturate or slightly degrade performance, suggesting that λ should be chosen in a moderate range.

Loss variant analysis. The proposed trial-consistency constraint aligns window-level predictive distributions within each trial. To test whether the effect depends on a particular divergence, we evaluate three variants under the same protocol: (i) KL divergence to the trial-mean prediction (**KL**, default), (ii) Jensen–Shannon divergence to the trial-mean prediction (**JS**), and (iii) a variance penalty(**Var**) that discourages within-trial dispersion of predicted class probabilities. All other settings are kept identical, and we report performance (accuracy) as well as reliability metrics (TC, TV) for comparison. The results are shown in Figure 2. In summary, **KL** performs best overall, yielding the strongest gains in both accuracy and trial consistency (TC) compared with JS-to-Mean and the variance penalty, while maintaining comparable or lower validation variation (TV). We therefore adopt KL-to-Mean as the default instantiation in all main experiments.

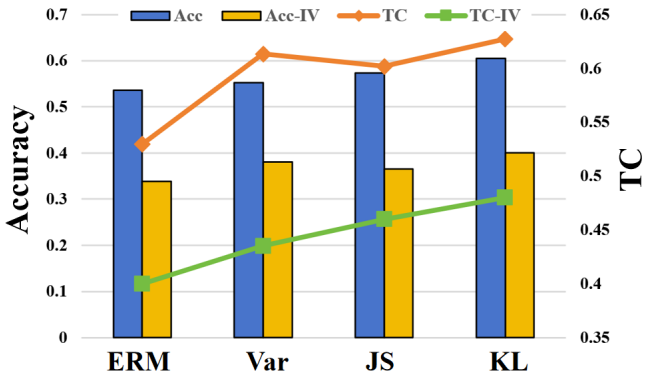

Figure 2: Effect of different trial-consistency regularizers using EEGNet. Bars and lines with "-IV" suffix: results on SEED-IV; bars and lines without suffix: results on SEED.

Negative control. To test whether the benefit of TrialCons depends on *genuine* within-trial coherence (rather than a generic regularization/smoothing effect), we run a shuffle control. Specifically, during training we randomly permute trial IDs across windows, thereby destroying the correspondence between windows and their original trial while leaving the model, loss form, and optimization settings unchanged. This control preserves the existence of grouped regularization but removes the intended semantic meaning of a trial as repeated measurements from the same stimulus segment. As shown in Figure 3, under the shuffled grouping, classification performance consistently degrades compared to using true trial IDs, supporting the interpretation that TrialCons is effective because it aligns predictions across *true* repeated measurements rather than merely smoothing outputs in an arbitrary grouping.

Generic smoothing baselines. A remaining alternative is that TrialCons gains arise from generic prediction smooth-

Figure 3: Trial-ID shuffle control: classification performance comparison across three conditions: ERM , TrialCons with true trial IDs, and TrialCons with shuffled trial IDs.

Table 4: Generic smoothing baselines using EEGNet.

Setting(on SEED)	Acc (%)↑	TC↑	TV↓
ERM	53.7±0.7	0.51±0.03	0.183±0.022
Label Smoothing	54.0±0.6	0.45±0.02	0.170±0.013
Mean Teacher	54.6±0.6	0.54±0.02	0.173±0.019
TrialCons	**60.3±0.7**	**0.63±0.01**	**0.150±0.020**

ing rather than trial-defined repeated-measurement structure. We therefore compare against two training-only regularizers that use no trial metadata: Label Smoothing (LS) and Mean Teacher (MT). All methods follow the same cross-session within-subject protocol and use the same backbone and training pipeline as ERM; neither LS nor MT uses trial IDs.

Label Smoothing (LS). We train with cross-entropy on softened targets $\tilde{y}_k = (1-\alpha)\mathbb{I}[k = y] + \alpha/K$ (with K=3 for SEED and K=4 for SEED-IV); α is tuned on validation.

Mean Teacher (MT). We maintain an EMA teacher θ' and add a consistency term between teacher and student predictions: $\mathcal{L}_{\text{mt}} = \mathcal{L}_{\text{ce}} + \lambda_{\text{mt}} D_{\text{KL}}(\mathbf{p}_{\theta'} \| \mathbf{p}_{\theta})$, where θ' is updated with momentum β; β and λ_{mt} are tuned on validation.

Results. Table 4 shows that generic smoothing alone cannot explain the gains of TrialCons. Both LS and MT yield only marginal accuracy changes relative to ERM and do not improve TC consistently (LS even reduces TC), while TrialCons delivers a large accuracy jump together with a substantial increase in trial consistency and lower validation variation. This pattern supports our claim that TrialCons benefits from exploiting trial-defined repeated-measurement structure rather than from generic regularization effects.

Robustness to Additive Gaussian Noise

Model robustness is evaluated via a controlled inference-time perturbation test. Each test window is first normalized using the same preprocessing as in the main experiments. Then, additive Gaussian noise is injected according to

$$\tilde{\mathbf{x}} = \mathbf{x} + \alpha\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{14}$$

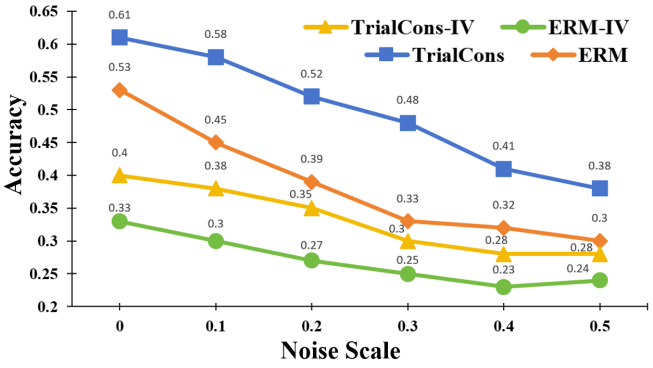

Figure 4: Noise robustness under additive Gaussian perturbations using EEGNet.Lines with "-IV" suffix: results on SEED-IV; Lines without suffix: results on SEED.

Interpreting degradation under different clean accuracies. Since TrialCons typically achieves higher clean accuracy at $\alpha = 0$, absolute accuracy drops are not directly comparable across methods. We therefore interpret robustness relative to the chance level, which is $1/3$ for SEED and $1/4$ for SEED-IV. Figure 4 shows ERM often reaches chance under moderate noise, whereas TrialCons maintains accuracy above chance over a wider range of α. This supports the view that TrialCons reduces sensitivity to task-irrelevant perturbations.

Computational Overhead

TrialCons adds no extra forward passes and does not modify the backbone. Its overhead comes only from (i) computing trial-wise mean probabilities within each mini-batch and (ii) evaluating the divergence terms.

We measure wall-clock overhead as average step time using synchronized CUDA timing after a warm-up phase, and report mean±std over 5 seeds. Under the default setting, TrialCons increases step time by **0.3% ± 0.02%** (always < 1%), confirming that the cost is negligible relative to backbone computation.

Conclusion

Trial-defined repeated measurements are intrinsic to stimulus-driven EEG emotion recognition, yet standard methods commonly treat windows as i.i.d., which can yield inconsistent within-trial predictions and unstable validation dynamics. This work introduces TrialCons, a loss function that uses training-set trial IDs to align window-level predictive distributions within each trial, without changing the backbone or the test-time interface. Experiments follow a cross-session protocol, with models instantiated by three standard backbones and evaluated under non-overlap and 50% overlap windowing. Reliability is assessed by trial consistency and validation total variation, together with test accuracy. Across SEED and SEED-IV, TrialCons consistently reduces epoch-to-epoch validation variability, and delivers stable accuracy gains, indicating smoother optimization and improved generalization.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0118001), the National Natural Science Foundation of China (Grant Nos. U22A2022, 62332017), the Beijing Nova Program (Grant No. 20250484951), and was conducted in part at the MOE Key Laboratory of Advanced Materials and Devices for Post-Moore Chips, the Beijing Key Laboratory of Big Data Innovation and Application for Skeletal Health Medical Care, and the Beijing Advanced Innovation Center for Materials Genome Engineering.

References

Abdul, A., Chen, J., Liao, H., & Chang, S. (2018). An emotion-aware personalized music recommendation system using a convolutional neural networks approach. *Applied Sciences*, *8*(7), 1103. https://doi.org/10.3390/app8071103

Alghamdi, A. M., Ashraf, M. U., Bahaddad, A. A., Almarhabi, K. A., Al Shehri, W. A., & Daraz, A. (2025). Cross-subject eeg signals-based emotion recognition using contrastive learning. *Scientific Reports*, *15*(1), 28295. https://doi.org/10.1038/s41598-025-13289-5

Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Chai, X., Wang, Q., Zhao, Y., Liu, X., Bai, O., & Li, Y. (2016). Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition. *Computers in Biology and Medicine*, *79*, 205–214. https://doi.org/10.1016/j.compbiomed.2016.10.019

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning (ICML)*.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation [Domain-Adversarial Neural Networks (DANN)]. *International Conference on Machine Learning (ICML)*.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*, 665–673. https://doi.org/10.1038/s42256-020-00257-z

Laine, S., & Aila, T. (2017). Temporal ensembling for semi-supervised learning. *International Conference on Learning Representations (ICLR)*.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *15*(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., Zhao, Z., Kumar, N., & Marttinen, P. (2023). Eeg based emotion recognition: A tutorial and review. *ACM Computing Surveys*, *55*(4), 1–57. https://doi.org/10.1145/3524499

Miyato, T., Maeda, S., Koyama, M., & Ishii, S. (2019). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(8), 1979–1993. https://doi.org/10.1109/TPAMI.2018.2858821

Picard, R. W. (1997). *Affective computing*. MIT Press.

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*.

Schlicher, M., Li, Y., Murthy, S. M. K., Sun, Q., & Schuller, B. W. (2025). Emotionally adaptive support: A narrative review of affective computing for mental health. *Frontiers in Digital Health*, *7*, 1657031. https://doi.org/10.3389/fdgth.2025.1657031

Song, Y., Zheng, Q., Liu, B., & Gao, X. (2023). EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 710–719. https://doi.org/10.1109/TNSRE.2022.3230250

Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation [arXiv:1607.01719]. *ECCV Workshops*.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, *30*.

Wang, F., Zhong, S., Peng, J., Jiang, J., & Liu, Y. (2018). Data augmentation for eeg-based emotion recognition with deep convolutional neural networks. *International Conference on Multimedia Modeling (MMM)*, *10705*, 82–93. https://doi.org/10.1007/978-3-319-73600-6_8

Zhao, W., Jiang, X., Zhang, B., Xiao, S., & Weng, S. (2024). CTNet: A convolutional transformer network for EEG-based motor imagery classification. *Scientific Reports*, *14*, 20237. https://doi.org/10.1038/s41598-024-71118-7

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., & Cichocki, A. (2019). EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, *49*(3), 1110–1122. https://doi.org/10.1109/TCYB.2018.2797176

Zheng, W.-L., & Lu, B.-L. (2015). 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. https://doi.org/10.1109/TAMD.2015.2431497