Improve Auditory Perturbation RSVP EEG Signal Decoding By Dual-view Backbone Based Dual-Stream Knowledge Distillation Network

Boxun Fu (sean2100@foxmail.com)
Jiachen Zhang (owl@stu.xidian.edu.cn)
Fu Li* (fuli@mail.xidian.edu.cn)
Xi Zhang (zxx@stu.xidian.edu.cn)
Junkai Li (junkaili@stu.xidian.edu.cn)
Xiaohui Cai (cxh@stu.xidian.edu.cn)
School of Artificial Intelligence, Xidian University, Xi'an, China

Abstract

Rapid Serial Visual Presentation (RSVP) is a critical cognitive paradigm for investigating visual attention dynamics and developing high-speed Brain-Computer Interfaces (BCIs). However, real-world deployment faces challenges from environmental noise, particularly cross-modal auditory perturbations that degrade neural decoding. This work proposes the Time-Frequency Dual-View Network (TFDV-Net), which fuses time-domain and frequency-domain features via a global-local interaction backbone to capture weak ERP signals. Furthermore, a Dual-Stream Knowledge Distillation strategy is introduced to enhance robustness, enforcing the model to reconstruct task-invariant cognitive patterns from noise-corrupted signals by learning from a "clean-state" teacher. The proposed method achieves a state-of-the-art balanced accuracy of 83.02% on a dataset containing ecological noise. Notably, we find that semantic conversational noise induces a significantly larger performance drop compared to non-semantic urban noise. This result quantitatively validates the hypothesis that high-level semantic processing imposes a greater cognitive load on the visual attention system than low-level acoustic interference.

Keywords: Brain-Computer Interface; EEG; RSVP; auditory perturbation; knowledge distillation

Introduction

Rapid Serial Visual Presentation (RSVP) serves as a fundamental paradigm in cognitive neuroscience to probe the temporal limits of human visual attention (Raymond et al., 1992; Shapiro et al., 1997). In this task, the brain must process continuous visual stimuli within millisecond-level windows to detect targets, eliciting specific event-related potentials (ERPs) (Robbins et al., 2020) - such as the P300 - that index the allocation of attentional resources (Polich, 2007). While Brain-Computer Interfaces (BCIs) leveraging this mechanism hold significant potential (Alpert et al., 2014; Chen et al., 2024; Marathe et al., 2016), the underlying cognitive process is inherently fragile due to the finite capacity of neural resources. In real-world environments, the brain faces not only exogenous signal artifacts but also complex endogenous interference-specifically, cross-modal cognitive competition triggered by environmental auditory stimuli. According to theories of intermodal attention (Koelewijn et al., 2010; Marti et al., 2012), this competition siphons a portion of shared executive resources away from the visual modality. Consequently, task-related EEG components (Millán et al., 2008) are attenuated rather than abolished, exhibiting diminished amplitudes and compromised signal-to-noise ratios compared to noise-free conditions (Fisher et al., 2023; Kaplan & Jesse, 2019; Kranczioch & Thorne, 2015), which severely limits decoding performance.

Existing EEG noise suppression methods predominantly target exogenous interference, such as power line noise or muscle/ocular artifacts. These artifacts are typically characterized by distinct frequency bands or fixed spatial distributions, making them manageable via conventional filtering or independent component analysis (ICA) (Bigdely-Shamlo et al., 2015). However, external sounds, once processed and cognitively integrated by the human auditory system, trigger obligatory neural responses that manifest as endogenous components within the EEG signal. Unlike random environmental noise, this endogenous interference conceptually overlaps with the target brain activity in both time and frequency domains. Consequently, it cannot be effectively removed using standard preprocessing techniques designed for exogenous artifacts (Wang et al., 2016), making robust decoding of RSVP EEG signals under auditory perturbation a challenging cognitive computing issue (Akça et al., 2023; Zhou et al., 2024).

To address this problem, this work proposes the Time-Frequency Dual-View RSVP EEG Decoding Network (TFDV-Net). This network utilizes a dual-stream input of time-domain and frequency-domain features, employing a global-local feature extraction backbone guided by global context to capture diminished ERP signals in noisy environments (Hazarika et al., 1997). By integrating multi-view representations, the model can better distinguish between the target visual ERPs and the interfering auditory neural responses, thereby significantly enhancing the decoding capability of RSVP-based BCIs.

Building on this backbone, we introduce a Dual-Stream Knowledge Distillation training strategy to further strengthen the network's resilience to cognitive interference. This strategy leverages a "Teacher" model trained on clean data to provide a robust reference for the "Student" model exposed to noise. Specifically, relational distillation transfers the structural processing patterns inherent in clean data, guiding the student model to maintain consistent inter-feature relationships even under perturbation. Simultaneously, feature distillation aligns the student's feature distribution with that of the teacher, minimizing distributional

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

discrepancies. This synergy effectively constrains the student network to approximate the clean feature manifold, enabling it to filter out noise components and extract task-relevant representations.

To verify the effectiveness of the proposed method, we constructed an ecologically valid RSVP EEG dataset incorporating urban and conversational noise perturbations. We evaluated TFDV-Net against several baseline methods on this dataset. Experimental results demonstrate that TFDV-Net achieves a balanced accuracy of 84.54% in a noise-free environment. Crucially, under noisy conditions, the model maintains strong robustness-particularly after applying the DSKD strategy-achieving a balanced accuracy of 81.90%. These findings confirm the method's efficacy in mitigating the impact of cross-modal interference on EEG decoding.

Method

This work proposes a TFDV-Net. The TFDV-Net involves a dual-view CNN-Transformer feature extraction backbone to fully capture ERP features of RSVP EEG. Based on the backbone, a dual-stream knowledge distillation training strategy is designed to enhance the decoding performance of RSVP EEG signals under noisy environments. The details are introduced as follows.

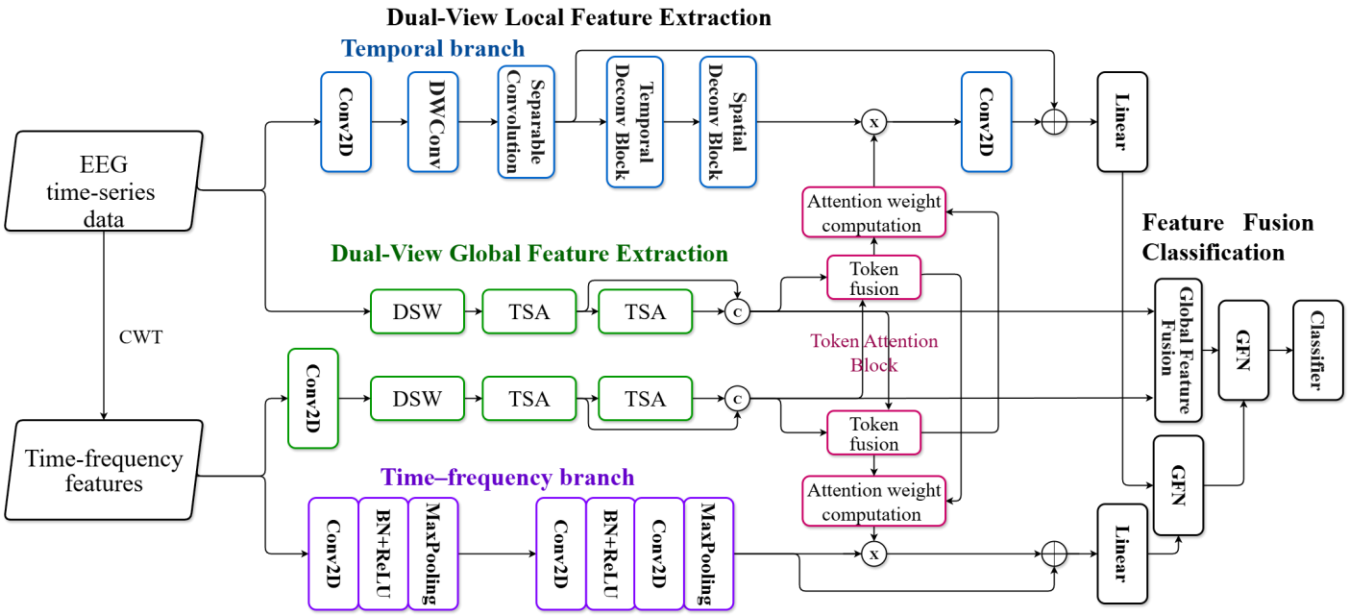

Figure 1: Time-Frequency Dual-View RSVP EEG Decoding Network.

Time-Frequency Dual-View RSVP EEG Decoding Network

The dual-view design integrates temporal waveform variations and time-frequency dynamic features across channels, enabling complementary multi-dimensional feature extraction. CNN and Transformer are combined to extract both local and global contextual features. The method takes preprocessed EEG time-series data and continuous wavelet transform (CWT) time-frequency features as dual-view inputs. The architecture consists of three main components: the Dual-View Global Feature Extraction, the Dual-View Local Feature Extraction, and the Feature Fusion Classification.

Dual-View Global Feature Extraction

The dual-view global feature extraction model dependencies of EEG signals across time and channels. It consists of the Dimension-Split Embedding (DSW) (Y. Zhang & Yan, 2023) Layer and the Two-Step Attention (TSA) (Y. Zhang & Yan, 2023) Layer, responsible for temporal segmentation embedding and cross-dimensional feature interaction, respectively.

a) In the **DSW layer**, the input EEG time-series are segmented along the temporal dimension and aligned with their corresponding channels. These segments are projected into the embedding space via linear projection and positional encoding, producing $Z \in \mathbb{R}^{L\times C\times D}$,where L is the number of temporal segments, C is the number of channels, and D represents the embedding dimension.

b) The **TSA layer** operates in two sequential steps: cross-time and cross-channel. In the cross-time step, a multi-head self-attention mechanism captures dependencies between different temporal segments within the same channel:

$$Z^{time}_{:,c} = LayerNorm(Z_{:,c} + MSA^{time}(Z_{:,c}, Z_{:,c}, Z_{:,c})), \quad (1)$$

where MSA^{time} denotes the cross-time multi-head self-attention, $Z^{time} \in \mathbb{R}^{L\times C\times D}$ represents features after temporal modeling.

In the inter-channel step, a routing vector $R \in \mathbb{R}^{L\times C\times D}$ serves as the query, interacting with channel-wise features to model inter-channel dependencies:

$$Z^{ch} = LayerNorm(Z^{time} + MSA^{ch}(Z^{time}, R, R)), \quad (2)$$

where MSA^{ch} denotes the inter-channel multi-head self-attention module, and the output $Z^{ch} \in \mathbb{R}^{L\times C\times D}$ integrates cross-dimensional relationships.

c) Finally, the outputs of the two TSA stages are concatenated and processed by a 2D convolution to reduce redundancy and enhance feature representation.

For time-series data input, the signals are processed by DSW followed by two TSA layers. For time-frequency input, the features are first compressed via 2D convolution with batch normalization and ReLU, and then encoded using the same DSW-TSA pipeline, yielding global time-domain and time–frequency-domain representations.

Dual-View Local Feature Extraction

The dual-view local feature extraction consists of temporal branch, time-frequency branch, and a token attention block, aiming to extract local features guided by global context.

a) The temporal branch employs a cascade of convolutional layers, depthwise separable convolutions, and transposed convolutions to sequentially capture temporal sequence features and inter-channel spatial features. The time–frequency branch is composed of multiple convolutional layers, batch normalization, ReLU activation, and pooling layers to extract multi-scale time-frequency features.

b) The Token Attention Block guides local feature extraction using global features. It is composed of a Token fusion layer and an attention weight computation layer. The Token fusion layer integrates temporal and time–frequency global features, calculating cross-modal correlations through bidirectional attention matrices A_{tf} and A_{ft}:

$$A_{ij=tf,ft} = \sigma_{softmax}(\frac{Q_i K_j^T}{\sqrt{d}}). \quad (3)$$

Key tokens are selected based on attention scores using a threshold strategy, and competitive fusion weights are assigned via Softmax, achieving weighted residual fusion of the two views.

The attention weight computation layer generates temporal and channel weights from the fused features. After compression and normalization, these weights are dynamically applied to the dual-view local feature extraction layers to suppress irrelevant components and enhance task-relevant signals. This cross-view interaction and dynamic weighting mechanism leverages global information to guide the focus on dual-view local features, thereby enhancing the network's ability to capture critical EEG details under low signal-to-noise ratio conditions.

Feature Fusion Classification

This Feature Fusion Classification integrates dual-view global and local features for final classification.

a) Local features are adaptively fused via a Gated Fusion Network (GFN) (Ren et al., 2018):

$$y_{CNN} = GFN(linear_1(X_{out1}), linear_2(X_{out2})), \quad (4)$$

where X_{out1} and X_{out2} are temporal and time-frequency local features.

b) Global features are concatenated, compressed via convolution and linear layers to obtain y_{global}.

c) Finally, local and global features are fused via GFN and fed to the classifier:

$$y_{out} = linear(GFN(y_{global}, y_{CNN})). \quad (5)$$

Dual-Stream Knowledge Distillation Training

To enable decoding of RSVP EEG under noisy perturbations, this work designs a Dual-Stream Knowledge Distillation Training framework, encompassing both the training strategy and the loss functions.

Training Strategy

To enhance the model's performance and robustness under noisy perturbations, we propose a dual-stream knowledge distillation approach. Since both clean and noisy data are derived from EEG signals elicited by the same sequence of stimulus positions (although the images may differ), the noisy EEG contains not only RSVP task-related ERP signals but also additional noise-induced components. The teacher model, trained on clean data, learns task-relevant ERP representations, which can guide the student model to disentangle noise while retaining core RSVP task-related features through knowledge distillation (Fu et al., 2024), thereby effectively improving the student model's ability to extract clean RSVP features from data under auditory noise perturbations.

The training procedure is conducted in two stages:

a) The teacher model is trained on clean data to extract stable and high-fidelity RSVP representations, capturing intrinsic data patterns and underlying structures to provide high-quality features for distillation.

b) The student model is trained on noisy data, leveraging both feature distillation and relation distillation to transfer knowledge from the teacher. Feature distillation minimizes the distribution discrepancy between teacher and student features, encouraging the student to approximate ideal representations and suppress noise components. Relation distillation preserves the inter-layer dependency patterns of features when both teacher and student process the same noisy inputs, further reinforcing feature decoupling. This process enhances the student's ability to extract RSVP-related features from noisy data while mitigating the learning of noise-induced auditory disturbance components.

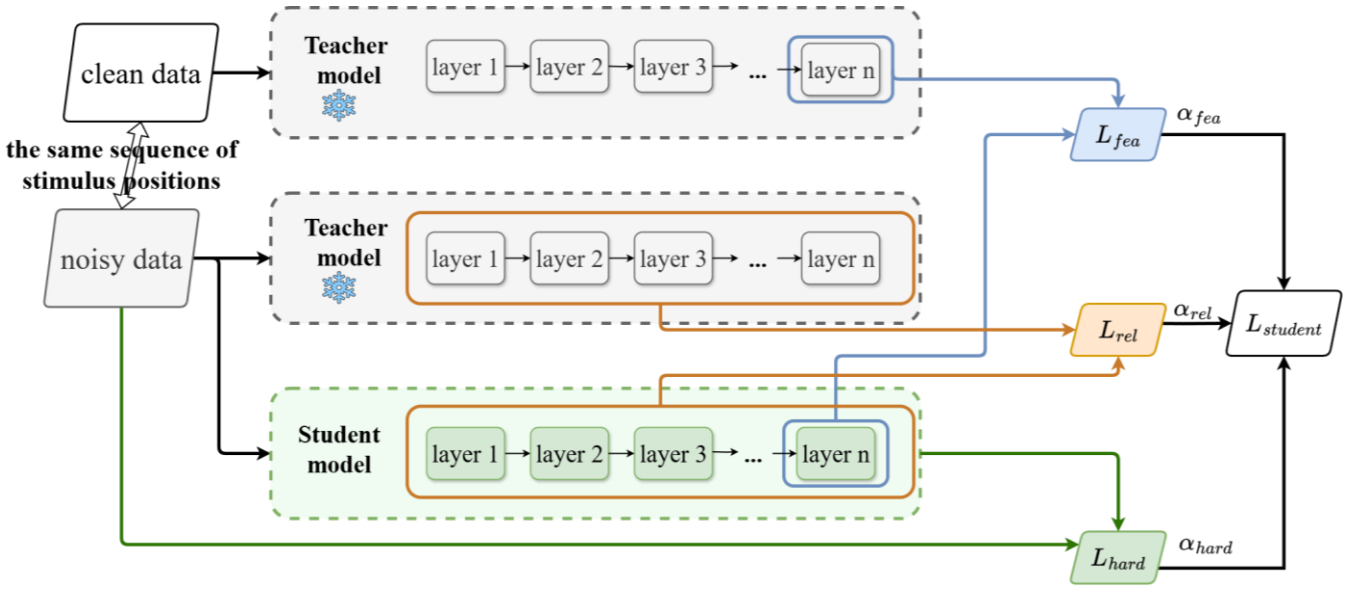

Figure 2: Dual-stream knowledge distillation framework, using clean data (data without noise perturbations) and noisy data (d ata under noise perturbation)

Loss Functions

Feature distillation is formulated using the KL divergence (Hinton et al., 2015) between teacher features and student feature, yielding the feature distillation loss L_{fea} .

Relation distillation measures the inter-layer relationships of the teacher and student models using Gramian matrices, whose correspondence is quantified via the FSP matrix (Yim et al., 2017). The relation distillation loss can be formulated as:

$$L_{rel}=\frac{1}{N}\sum_{x}\sum_{i=1}^{n}\lambda_i\left\|FSP(G_i^T(x)-G_i^S(x))\right\|_2^2, \qquad (6)$$

where $G_i^T(x)$ and $G_i^S(x)$ denote the Gramian-derived (Yim et al., 2017) inter-layer relations of the i -th feature pair for the teacher and student models, and λ_i represents the weighting factor for each layer. The FSP matrix characterizes the propagation dynamics and inter-layer dependency patterns of feature representations.

The student model is ultimately trained by jointly optimizing the feature distillation loss L_{fea} , relation distillation loss L_{rel} , and the standard hard label loss L_{hard} :

$$\begin{aligned} &L_{student}=\alpha_{rel}L_{rel}+\alpha_{fea}L_{fea}+\alpha_{hard}L_{hard},\\ &s.t.\alpha_{rel}+\alpha_{fea}+\alpha_{hard}=1, 0\le\alpha_{rel},\alpha_{fea},\alpha_{hard}\le 1 \end{aligned}. \qquad (7)$$

This strategy fully leverages idealized representations from clean data, ensuring effective decoupling of task-relevant ERP features from noise, and enables the student model to acquire stable and robust representations under challenging conditions. Ultimately, it enhances the student model's decoding performance of EEG signals in RSVP tasks under noisy perturbations.

Experiments

Experimental Setup

To evaluate the performance, effectiveness, and robustness of the proposed method for RSVP EEG decoding under auditory noise perturbations, we constructed an RSVP EEG experiment based on the standard RSVP paradigm with controlled auditory noise interference and collected a local dataset. The dataset consists of EEG recordings from nine healthy participants under three conditions: no noise (ideal environment), urban noise (non-semantic noise), and conversational noise (semantic noise) which were used to obtain clean data, urban-noise data, and conversational-noise data, respectively. EEG signals were recorded using a BioSemi ActiveTwo system (64 electrodes, reference electrodes placed below P1 and P2, sampling rate 1024 Hz, electrode impedance < 25 kΩ). Preprocessing involved 1–49 Hz bandpass filtering, downsampling to 256 Hz, and 1-second segmentation aligned to stimulus onset, yielding final dimensions of 64×256. Each participant contributed 512 samples balanced between target and non-target stimuli, which were partitioned into training, testing, and validation sets with a ratio of 7:2:1.

Four representative baselines were selected: a traditional machine learning method, HDCA (Parra et al., 2008); two classical deep learning approaches, EEGNet (Lawhern et al., 2018) and EEG-Inception (Santamaría-Vázquez et al., 2020); and a recent RSVP-specific model, TSFormer-SA (Li et al., 2024). All models were implemented in PyTorch and trained using cross-entropy loss with the Adam optimizer (learning rate 0.001) for 200 epochs. The model achieving the lowest validation loss was selected for final testing. Experiments were conducted on Ubuntu 22.04 LTS with an Intel® i7-10700KF CPU, NVIDIA RTX 4090 GPU, and 32 GB RAM.

Experimental Results

The Table 1 presents the balanced accuracy (BA) results of the proposed TFDV-Net and baseline methods under three experimental protocols and two types of noise perturbations:

A: training and testing on clean data;

B: training on clean data and testing on noisy data;

C: training by the proposed Dual-Stream Knowledge Distillation method and testing on noisy data.

The Protocol A reflects the performance of the backbone network, Protocol B evaluates its robustness against noise-contaminated data, and Protocol C demonstrates the effectiveness of the proposed training strategy in enhancing decoding performance on noisy data based on the backbone model.

Table 1: Experimental results of Balanced Accuracy (BA) for RSVP EEG decoding under auditory noise perturbation.

Model		Urban Noise (%)	Conversational Noise (%)	Mean (%)
HDCA	A	/	/	76.39±4.84
	B	62.43±2.67	59.96±2.72	61.20±2.76
	C	/	/	/
EEGNet	A	/	/	81.83±4.24
	B	71.74±5.25	69.01±5.22	70.41±7.61
	C	76.16±6.75	71.18±5.90	73.58±6.31
EEG-Inception	A	/	/	82.30±3.30
	B	72.77±4.37	70.23±4.29	71.16±5.60
	C	76.07±5.79	71.67±5.05	73.87±5.55
TSformer-SA	A	/	/	83.30±1.99
	B	75.38±3.53	73.23±2.52	73.47±4.01
	C	79.75±1.03	76.42±2.68	78.09±3.21
TFDV-Net	A	/	/	84.54±3.12
	B	79.07±4.75	77.60±3.45	78.34±4.59
	C	83.02±3.86	80.78±4.31	81.90±4.34

On average, TFDV-Net achieved a BA of 84.54% under Protocol A, outperforming HDCA (76.39%) by 8.15 %, EEGNet (81.83%) by 2.71 %, and TSformer-SA (83.30%) by 1.24%, verifying the superior decoding performance of the proposed backbone on standard RSVP EEG. Under Protocol B, TFDV-Net achieved 78.34%, the highest among all methods. Its performance degradation from A → B was −5.47%, smaller than EEGNet (−11.45%), EEG-Inception (−10.08%), and TSformer-SA (−8.99%), indicating stronger inherent robustness to noise perturbations. Under Protocol C, TFDV-Net reached 81.90% BA, with only −2.64% performance degradation from A → C, outperforming TSformer-SA (78.09%) by +3.81% and showing 2.57% less performance degradation from A→C. This demonstrates that the proposed dual-stream knowledge distillation effectively recovers performance under noisy conditions.

For both urban and conversation noise, all models degraded more severely under semantic noise, while TFDV-Net consistently achieved the best BA and exhibited the smallest decline across Protocols B and C, demonstrating its superior robustness and the effectiveness of the proposed training strategy.

Analysis and Discussions

Ablation Study

To systematically verify the effectiveness of TFDV-Net under complex acoustic environments, we conducted stepwise ablation experiments under Urban Noise and Conversational Noise conditions. Using Balanced Accuracy (BA) as the core metric, we evaluated the components in the order of the model's processing pipeline: the dual-view backbone (Time-Frequency Branch and Temporal Branch), the feature enhancement module (Token Attention Block), and the final fusion strategy (Feature Fusion Classification). The detailed experimental results are presented in Table 2.

Table 2: Ablation Study Experimental Results of Balanced Accuracy (BA) for RSVP EEG Decoding with TFDV-Net Components under Different Auditory Noise Perturbations.

Model Variants	Urban Noise (%)	Conversational Noise (%)
Proposed Full Model	79.07±4.75	77.60±3.45
w/o Time-Frequency Branch	75.84±6.33	73.13±6.15
w/o Temporal Branch	62.38±7.71	64.98±4.23
w/o Token Attention Block	75.35±5.21	73.34±5.17
w/o Feature Fusion Classification	74.79±2.88	71.45±5.55

We first evaluated the input branches. The Time-Frequency Branch is responsible for capturing local spatial-frequency patterns. As shown in Table I, removing this branch reduced the BA to 75.84% and 73.13% under Urban and Conversational noise, respectively. Notably, the absence of this branch significantly increased the standard deviation in the Urban scenario to ±6.33%, highlighting the critical role of time-frequency features in maintaining performance stability across different subjects. In contrast, the Temporal Branch played a decisive role. Its removal caused the most precipitous performance drop, with BA plummeting to 62.38% and 64.98% under the two noise conditions. This strongly demonstrates that the global temporal dynamics of EEG signals serve as the primary discriminative basis for RSVP target detection, and time-frequency features alone are insufficient to resist strong environmental noise interference.

Following feature extraction, the Token Attention Block is introduced to enhance key representations. Removing this module resulted in a BA decrease of 3.72% and 4.26% across the two noise environments. This indicates that by computing attention weights, the network can effectively filter high-value neural response patterns from contaminated signals and

suppress noise interference in intermediate features, thereby maintaining high discriminative accuracy under low signal-to-noise ratio conditions.

Finally, regarding the decision fusion strategy, we compared the Gated Fusion Network with traditional linear concatenation. Removing the Feature Fusion Classification (i.e., reverting to linear concatenation) degraded performance by 4.28% and 6.15% under Urban and Conversational noise, respectively. This confirms that at the decision stage, adaptively adjusting the contribution weights of dual-view features offers superior robustness compared to fixed concatenation. This gain is particularly evident under Conversational Noise, suggesting that dynamic fusion better handles complex semantic interference.

In summary, the Proposed Full Model achieves optimal performance (79.07% / 77.60%) with the lowest standard deviation across all conditions, leveraging synergistic optimization from front-end feature complementarity to back-end dynamic fusion.

Hyperparameter Sensitivity Analysis

In the dual-stream knowledge distillation framework, the total loss function is composed of the relation distillation loss (L_{rel}), feature distillation loss (L_{fea}), and hard label classification loss (L_{hard}). The weight coefficients α_{rel} , α_{fea} , and α_{hard} determine the trade-off between maintaining classification discriminability and transferring knowledge from the teacher model. To investigate the optimal configuration, we evaluated five different parameter combinations under Urban Noise and Conversational Noise conditions. The experimental results, reported as Mean ± Standard Deviation (Std), are summarized in Table 3.

Table 3: Comparison of Balanced Accuracy (BA) under Different Loss Function Coefficient Settings.

α_{rel}	α_{fea}	α_{hard}	Urban Noise (%)	Conversational Noise (%)
0.2	0.3	0.5	76.29 ± 2.82	73.69 ± 2.01
0.3	0.2	0.5	75.78 ± 1.62	72.59 ± 2.44
0.1	0.2	0.7	83.02 ± 3.86	80.78 ± 4.34
0.2	0.1	0.7	82.05 ± 3.31	80.28 ± 3.44
0.05	0.05	0.9	81.43 ± 4.14	77.75 ± 3.95

The results demonstrate that the allocation of weight coefficients obviously impacts both model performance and stability. When α_{hard} is set to 0.5, the model achieves the lowest performance across both datasets (76.29% and 75.78% under Urban Noise, 72.59% and 73.69% under Conversational Noise). Although the standard deviation is relatively low (2.82% and 1.62% under Urban Noise， 2.01% and 2.44% under Urban Noise), this primarily reflects a "low-performance stability," where the model fails to establish sufficiently clear decision boundaries to distinguish complex samples, and distillation knowledge alone cannot compensate for this deficiency.

Increasing α_{hard} to 0.7 leads to a substantial performance boost. At this level, the configuration prioritizing feature discrepancy constraint ($\alpha_{rel}=0.1,\alpha_{fea}=0.2$) slightly outperforms the one prioritizing relation constraint ($\alpha_{rel}=0.2,\alpha_{fea}=0.1$). While the standard deviation increases (reaching ±3.86% and ±4.34% under Urban and Conversational Noise), this reflects the variance in adaptability across subjects, but the overall performance gain outweighs the stability risk. However, when α_{hard} is further increased to 0.9, performance declines, and the standard deviations become ±4.14% and ±3.95% under Urban and Conversational Noise, respectively. indicating increased sensitivity to noise and reduced stability due to the excessive compression of distillation weights.

In summary, $\alpha_{rel}=0.1,\alpha_{fea}=0.2,\alpha_{hard}=0.7$ is identified as the optimal parameter setting. Under this configuration, the model achieves the highest Balanced Accuracy of 83.02% and 80.78% under Urban and Conversational Noise, respectively, demonstrating that a moderate classification dominance combined with feature-level distillation constraints effectively enhances robustness in noisy environments.

Conclusion

This work addresses the challenge of robustly decoding RSVP EEG signals under cross-modal auditory interference by proposing TFDV-Net. Unlike traditional denoising methods, TFDV-Net utilizes a global-local interaction backbone to extract comprehensive features from both time and frequency domains. To further enhance noise resilience, we introduce a Dual-Stream Knowledge Distillation strategy. By leveraging a teacher model trained on clean data as a structural reference, this strategy employs relational distillation to enforce consistent feature topology and feature distillation to minimize distributional discrepancies. This synergy allows the student model to effectively suppress noise-related patterns and reconstruct clean cognitive representations. Experimental results on an ecologically valid dataset demonstrate that TFDV-Net achieves a balanced accuracy of 83.02%, significantly outperforming baselines. Crucially, our analysis reveals that semantic conversational noise degrades decoding performance more severely than non-semantic urban noise, providing quantitative evidence for the high cognitive cost of linguistic interference. Future work will focus on adapting this framework to application-oriented BCIs operating in naturalistic, real-world acoustic environments.

References

Akça, M., Vuoskoski, J. K., Laeng, B., & Bishop, L. (2023). Recognition of brief sounds in rapid serial auditory presentation. *PLOS ONE*, 18(4), e0284396.

Alpert, G. F., Manor, R., Spanier, A. B., Deouell, L. Y., & Geva, A. B. (2014). Spatiotemporal representations of rapid visual target detection: A single-trial EEG classification algorithm. *IEEE Transactions on Biomedical Engineering, 61*(8), 2290–2303.

Biasiucci, A., Franceschiello, B., & Murray, M. M. (2019). Electroencephalography. *Current Biology, 29*(3), R80–R85.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*, 16.

Chen, L., Fan, Y., Gao, Y., Qiu, S., Wei, W., & He, H. (2024). EEG-FRM: A neural network based familiar and unfamiliar face EEG recognition method. *Cognitive Neurodynamics, 18*(2), 357–370.

Fisher, V., Dean, C. L., Nave, C. S., Parkins, E. V., Kerkhoff, W. G., & Kwakye, L. D. (2023). Increases in sensory noise predict attentional disruptions to audiovisual speech perception. *Frontiers in Human Neuroscience, 16*, 1077724.

Fu, Y., Huang, B., Wen, Y., & Zhang, P. (2024). FDR-MSA: Enhancing multimodal sentiment analysis through feature disentanglement and reconstruction. *Knowledge-Based Systems, 297*, 111965.

Gherri, E., & Eimer, M. (2011). Active listening impairs visual perception and selectivity: An ERP study of audio-visual dual-task costs. *Journal of Cognitive Neuroscience*, *23*(4), 832–844.

Hazarika, N., Chen, J. Z., Tsoi, A. C., & Sergejew, A. (1997). Classification of EEG signals using the wavelet transform. *Signal Processing, 59*(1), 61–72.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv*. https://doi.org/10.48550/arXiv.1503.02531

Kaplan, E., & Jesse, A. (2019). Fixating the eyes of a speaker provides sufficient visual information to modulate early auditory processing. *Biological Psychology*, 146, 107724.

Koelewijn, T., Bronkhorst, A. W., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica*, *134*(3), 372–384.

Kranczioch, C., & Thorne, J. D. (2015). The beneficial effects of sounds on attentional blink performance: An ERP study. *NeuroImage*, 117, 429–438.

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.

Li, X., Wei, W., Qiu, S., & He, H. (2024). A temporal–spectral fusion transformer with subject-specific adapter for enhancing RSVP-BCI decoding. *Neural Networks*, 181, 106844.

Marathe, A. R., Lawhern, V. J., Wu, D., Slayback, D., & Lance, B. J. (2016). Improved neural signal classification in a rapid serial visual presentation task using active learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 24*(3), 333–343.

Marti, S., Sigman, M., & Dehaene, S. (2012). A cortical network for reflective cognitive control. *Science, 335*(6075), 1488–1493.

Millán, J. D. R., Ferrez, P. W., Galán, F., Lew, E., & Chavarriaga, R. (2008). Non-invasive brain-machine interaction. *International Journal of Pattern Recognition and Artificial Intelligence, 22*(05), 959–972.

Parra, L. C., Christoforou, C., Gerson, A. D., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M. G., & Sajda, P. (2008). Spatiotemporal linear decoding of brain state. *IEEE Signal Processing Magazine, 25*(1), 107–115.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849.

Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., & Yang, M.-H. (2018). Gated fusion network for single image dehazing. I*n Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3253–3261).

Robbins, K. A., Touryan, J., Mullen, T., Kothe, C., & Bigdely-Shamlo, N. (2020). How sensitive are EEG results to preprocessing methods: A benchmarking study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28*(5), 1081–1090.

Santamaría-Vázquez, E., Martínez-Cagigal, V., Vaquerizo-Villar, F., & Hornero, R. (2020). EEG-Inception: A novel deep convolutional neural network for assistive ERP-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28*(12), 2773–2782.

Schembri, P., Pelc, M., & Ma, J. (2020). The effect that auditory distractions have on a visual P300 speller while utilizing low-cost off-the-shelf equipment. *Computers*, *9*(3), 68.

Shapiro, K. L., Raymond, J. E., & Arnell, K. M. (1997). The attentional blink. *Trends in Cognitive Sciences*, *1*(8), 291–296.

Spence, C. (2010). Crossmodal spatial attention. *Annals of the New York Academy of Sciences*, *1191*(1), 182–200.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410.

Wang, N. Z., Healy, G., Smeaton, A. F., & Ward, T. E. (2016). An investigation of triggering approaches for the rapid

serial visual presentation paradigm in brain computer interfacing. *Dublin City University Open Access Institutional Repository*.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455.

Yim, J., Joo, D., Bae, J., & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 107–115).

Zhang, S., Wang, Y., Zhang, L., & Gao, X. (2020). A benchmark dataset for RSVP-based brain–computer interfaces. *Frontiers in Neuroscience, 14*, 568000.

Zhang, Y., & Yan, J. (2023). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhou, Q., Zhang, Q., Wang, B., Yang, Y., Yuan, Z., Li, S., Zhao, Y., Zhu, Y., Gao, Z., Zhou, J., & Wang, C. (2024). RSVP-based BCI for inconspicuous targets: Detection, localization, and modulation of attention. *Journal of Neural Engineering, 21*(4), 046046.