Expert Systems With Applications 288 (2025) 128189

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Brain-inspired deep learning model for EEG-based low-quality video target detection with phased encoding and aligned fusion

Dehao Wang [a], Jianting Shi [b], Manyu Liu [b], Wenao Han [b], Luzheng Bi [b], Weijie Fei [b,*]

[a] School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, 100081, China
[b] School of Mechanical Engineering, Beijing Institute of Technology, Beijing, 100081, China

ARTICLE INFO

Keywords:
Brain-computer interface
Brain-inspired
Electroencephalogram
Low-quality video target detection

ABSTRACT

Brain-computer interface (BCI) technologies for video target detection hold great promise across various applications. However, existing algorithms exhibit limited performance in electroencephalogram (EEG) decoding for target detection in low-quality videos. In this paper, to address the limitation, we propose a novel brain-inspired deep learning model that incorporates EEG phased encoding and feature-aligned fusion. We first divide the EEG segments into pre-phase and post-phase, and extract the corresponding compressed temporal features using a novel phased encoder, which is based on multi-scale convolution and attention mechanisms. Subsequently, to capture the full-phase brain response, we align and integrate the features from both phases and extract global temporal features for classification. The proposed model is grounded in our time- and frequency-domain neural analysis, which identifies three critical phases of the brain's response during low-quality video target detection: early target recognition, later target spatial tracking, and sustained attention throughout the entire phase. EEG datasets, with and without ICA-based artifact removal, were used for cross-subject training and evaluation, with the proposed model consistently outperforming baselines. Pseudo-online tests confirmed real-time performance, and additional experiments with cognitively distracted participants further demonstrated the model's robustness. This work addresses a significant gap in low-quality video target detection algorithms and advances brain-inspired EEG classification by combining principles of neuroscience with artificial intelligence techniques. Our code is available at: https://github.com/Wonder-How/PSAFNet.

1. Introduction

Video target detection has critical applications across various fields, such as pedestrian and vehicle detection in autonomous driving (Feng, Harakeh, Waslander, & Dietmayer, 2021), defect detection in industrial production (Li, Zhang, Wang, Yang, & Deng, 2022a), and security (Yun et al., 2022), ecological monitoring (Lyu et al., 2022), and disaster response (Alawad, Halima, & Aziz, 2023) using unmanned aerial vehicles (UAVs). However, many videos, particularly aerial footage, suffer from low quality due to environmental interference, low resolution, and motion instability (Garvanov, Garvanova, Ivanov, Chikurtev, & Chikurteva, 2024; Wang et al., 2023). Addressing the challenge of target detection in these low-quality videos has become a prominent issue.

The target detection problem in UAV-captured videos is typically addressed through computer vision or manual detection methods. However, two major challenges remain despite advancements in deep learning: (1) the aerial perspective, affected by factors such as weather, frequently leads to obstructed, fragmented, or small targets, which greatly reduces detection accuracy (Fang, Zhang, Zheng, & Chen, 2023); and (2) the inherent uncertainty in scenarios like military reconnaissance or disaster response often lacks prior information about targets, complicating detection and limiting the effectiveness of deep learning models in few-shot learning tasks (Liu et al., 2019).

In contrast, the human brain excels at target detection in low-quality videos due to its reasoning and adaptability. For challenge (1), the brain can infer occluded or blurred targets based on background context and prior experiences (Bar, 2007; Mansfield, 2024). For challenge (2), the brain's few-shot learning ability allows it to recognize objects by combining experience with the current task (Pourpanah et al., 2022), eliminating the need for extensive pretraining data (Geirhos et al., 2018). However, in low-quality scenarios, some targets are difficult to identify with the naked eye, and manual reporting introduces delays, making

* Corresponding author: Weijie Fei School of Mechanical Engineering, Beijing Institute of Technology, Beijing, 100081, China.
E-mail addresses: wonderhow@bit.edu.cn (D. Wang), jat_shi@bit.edu.cn (J. Shi), 3120230395@bit.edu.cn (M. Liu), 3120240410@bit.edu.cn (W. Han), bhxblz@bit.edu.cn (L. Bi), bitfwj@bit.edu.cn (W. Fei).

https://doi.org/10.1016/j.eswa.2025.128189
Received 13 January 2025; Received in revised form 12 May 2025; Accepted 14 May 2025
Available online 20 May 2025

it unsuitable for real-time detection in urgent situations. Furthermore, manual reporting is often impractical in real-world multitasking environments.

In recent years, the rapid development and application of brain-computer interface (BCI) technology have provided a promising solution for real-time automated target detection in complex scenarios. As a highly promising next-generation technology, BCI enables interaction between the brain and external devices by capturing neural activity (Wang et al., 2024, 2022; Xia et al., 2024). Electroencephalography (EEG), known for its non-invasiveness, high temporal resolution, and affordability, has become widely used in BCI, particularly for target detection. In cases where potential targets cannot be identified visually, EEG can capture participants' unconscious neural responses. Additionally, EEG-based target recognition promotes system automation, allowing users to focus on other critical tasks simultaneously.

Currently, research on using BCI for low-quality video target detection remains limited. Most previous EEG-based target detection studies have focused on images, primarily following the Rapid Serial Visual Presentation (RSVP) paradigm. This paradigm presents images to the subject at a high rate, using the Event-Related Potentials (ERP) elicited by the subject to detect the target. Sajda, Gerson, and Parra (2003) were the first to apply the RSVP task to high-throughput image target detection. Their experiments showed that EEG classification often outperformed the traditional manual button pressing method in detecting targets. Various EEG decoding algorithms have been employed in RSVP tasks, such as Hierarchical Discriminant Component Analysis (Gerson, Parra, & Sajda, 2006), Spatial Filtering xDAWN (Rivet, Souloumiac, Attina, & Gibert, 2009), and Minimum Distance to Riemannian Mean (Barachant & Congedo, 2014). However, the classification accuracy of these methods still requires improvement.

Deep learning methods have been widely applied in EEG-based target detection. Lawhern et al. (2018) proposed EEGNet, using depthwise separable convolutions to analyze EEG features. Zang, Lin, Liu, and Gao (2021) developed PLNet, leveraging the phase-locked characteristics of ERP signals to extract features across different time windows. Li, Wei, Qiu, and He (2022b) proposed the Temporal-Frequency Fusion Transformer (TFF-Former), a multi-view fusion framework designed to capture shared temporal-frequency features across subjects. Yuan et al. (2024) employed pyramid squeeze attention for single-trial RSVP task. However, most existing BCI target detection algorithms are designed for image data, with relatively little research focusing on video-based target detection. Adapting these algorithms to the video target detection paradigm presents a significant challenge.

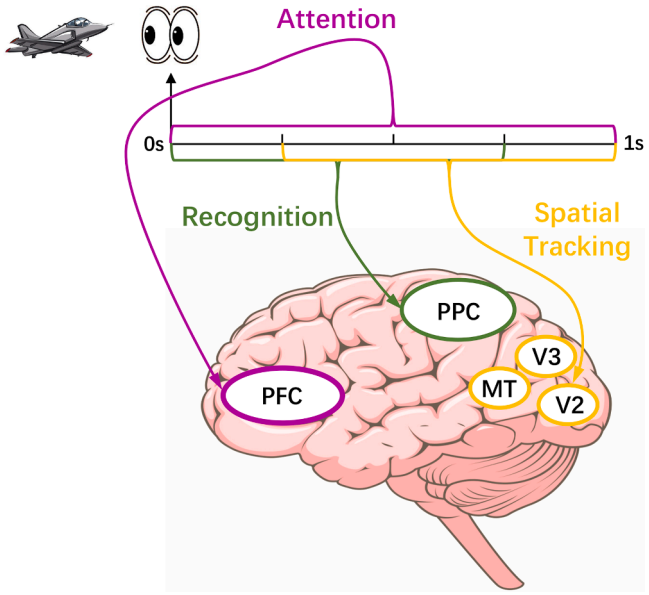

Fig. 1. The brain mechanisms in low-quality video target detection.

For low-quality video target detection, our previous work utilized FRP (Fixation-Related Potential) instead of ERP as the feature segment representing target detection (Shi, Bi, Xu, Feleke, & Fei, 2024). This approach partially mitigated the asynchrony between target appearance and participant recognition. However, there remains substantial research potential in designing algorithms to classify targets based on the extracted FRP segments.

Designing algorithmic models inspired by the brain mechanisms of video target detection is a promising approach. There have been some studies on brain-inspired algorithms for EEG. Inspired by the brain's capability to handle uncertain, noisy, and incomplete information, Type-2 fuzzy logic offers a robust framework for EEG signal decoding and cognitive modeling (Nguyen, Khosravi, Creighton, & Nahavandi, 2015; Rahmani, Mohajelin, Khaleghi, Sheykhivand, & Danishvar, 2024). Spiking Neural Networks (SNNs) transform EEG signals into spike sequences to mimic biological neural processing (Choi, 2024). Hierarchical Temporal Memory (HTM), inspired by the structure of the cerebral cortex, has proven effective in capturing temporal patterns in EEG data (Struye & Latré, 2020). Additionally, Hebbian learning, rooted in neural self-organization principles, offers further potential for EEG signal modeling (Uleru, Hulea, & Manta, 2022). Wendling et al. (2024) proposed brain-inspired computational models of the human cortex, structured at the cellular, assembly, and whole-brain levels, to support the diagnosis of epilepsy.

For visual brain-computer interfaces, Related research has shown that the brain processes visual tasks in stages (Song et al., 2021). Inspired by this, Lu, Zeng, Zhang, Yan, and Tong (2022b) introduced SAST-GCN, which segments EEG data into three non-overlapping phases, extracts features using graph convolution, concatenates them temporally, and classifies the results with a convolutional network, achieving 90.55 % accuracy. However, their method focuses only on high-quality videos, differing significantly from the low-quality scenarios in practical applications. It does not incorporate complex brain processes such as spatial tracking or long-term attention, making it unsuitable for low-quality video target detection tasks discussed above.

Unlike the brief experimental paradigm of RSVP or high-quality video tasks, low-quality video target detection involves complex, multi-phase brain mechanisms. Neural representation analysis categorizes this task into three key phases, as illustrated in Fig. 1. The early phase involves the surprise response induced by the appearance of the target and the recognition and evaluation of task-related targets. These two neural responses correspond to the components P3a and P3b of P300, occurring approximately within the early phase after the observer sees the target (Polich, 2007). The primary brain region involved is the posterior parietal cortex (PPC), marked in green in the figure. The later phase occurs after the target is detected, when the movement of the video target induces a brain response related to visual spatial tracking. This response is primarily localized in the visual areas of brain, particularly in V2, V3, and the MT area responsible for motion perception of objects (Born & Bradley, 2005; Jahn, Wendt, Lotze, Papenmeier, & Huff, 2012; Kaas, 2003). The response occurs roughly in later phase after target detection, as indicated in yellow in the figure. The full phase involves sustained attention and cognitive responses related to target perception throughout the entire detection process. This phase is primarily associated with the prefrontal cortex (PFC) (Knight, 1994; Martinez-Trujillo, 2022), with responses occurring across the 0–1 second time span, marked in purple in the figure. These neural responses across phases illustrate how the brain coordinates the processing of a target's appearance, movement, and cognitive interpretation during low-quality video target detection.

Inspired by the brain mechanisms involved in low-quality video target detection, we propose a deep learning model based on a segmented temporal encoder and aligned fusion. For the early phase (recognition)

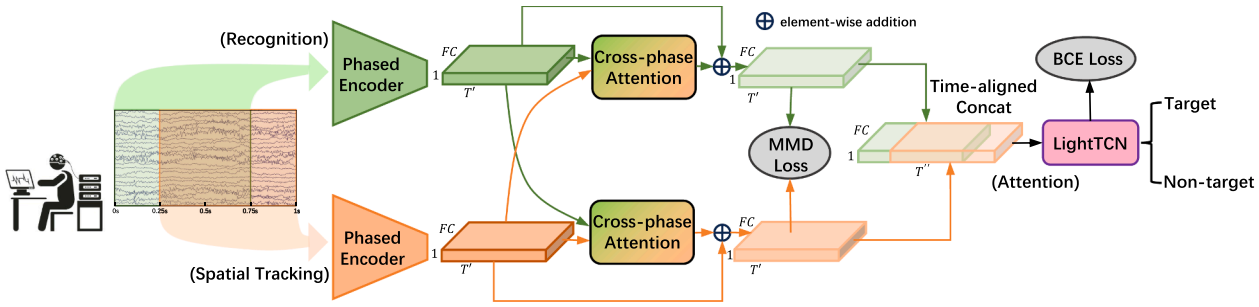

Fig. 2. The overall architecture of our brain-inspired video target detection model. MMD: Maximum Mean Discrepancy; BCE: Binary Cross Entropy; LightTCN: Light temporal convolutional network. T represents the time dimension, and F represents the number of feature layers. Processed through the Phased Encoder, the spatial channel dimension remains 1. The Phased Encoder is shown in Fig. 3, the Cross-phase Attention is shown in Fig. 4, and the LightTCN is shown in Fig. 5.

and later phase (spatial tracking) of video target detection, a 1-second time segment is divided into two overlapping phases: the first 0.75 second and the last 0.75 second. Each phase is processed through a proposed Phased Encoder to extract temporal and spatial features. Next, to capture the full phase response of the brain (attention), we align and temporally match the two extracted features, followed by concatenation. A temporal network is then used to extract the full-phase temporal features, which are employed to determine whether the EEG segment corresponds to target detection.

The contribution of this paper is as follows: We propose a novel brain-inspired deep learning model that incorporates EEG phased encoding and feature-aligned fusion for EEG-based low-quality video target detection. Through time- and frequency-domain analysis of EEG signals, we find that the brain's response is divided into early target recognition, later spatial tracking, and full-phase attention concentration, corresponding to the model's early and later phase encoding and feature alignment. By using EEG data with and without ICA, we simulated pure brain signals and artifact-mixed signals. Additionally, cognitive distraction experiments were conducted to assess the model's performance in multitasking scenarios. Across various conditions, our model consistently outperformed baseline models, which highlights the model's superior accuracy and robust generalization capabilities across diverse scenarios and conditions.

2. Method

In this study, inspired by the brain mechanisms involved in video target recognition, we propose a deep learning model with phased alignment and fusion. For a 1-second EEG segment, we first divide it into two overlapping parts: the first 0.75 second (recognition) and the last 0.75 second (spatial tracking). Each segment is then processed by a Phased Encoder to extract compressed temporal features. The encoder consists of four block, which are detailed in Section 2.1. Subsequently, to capture global features (attention), the temporal features from both phases are aligned in high-dimensional space and temporal relations. These features are fused through cross-attention mechanisms and a temporal network, ultimately yielding the classification result, as described in Section 2.2. The overall model architecture is illustrated in Fig. 2.

2.1. Phased encoder

To effectively capture the spatial-temporal characteristics of different phased EEG signals, we propose a novel encoder designed to extract relevant features from segmented EEG data, as illustrated in Fig. 3. The Phased Encoder leverages a multi-scale temporal convolution to process temporal dynamics in the EEG signals. It incorporates spatial attention mechanisms to focus on critical regions of interest across different brain areas, while also optimizing the integration of different feature channels. Through the combination of these strategies, the Phased Encoder efficiently compresses and extracts meaningful spatial and representations from the EEG data, which are then used for further alignment and fusion.

2.1.1. Multi-scale temporal convolution

The model takes a segmented EEG fragment matrix, $X \in \mathbb{R}^{C\times T}$, as input, where C represents the number of channels and T denotes the number of time points. The EEG is processed in parallel through three temporal convolution modules, each with a kernel size of 1 in the spatial domain and temporal window sizes of 32, 64, and 96, respectively, with a stride of 1. Each temporal convolution module outputs $FC_1/3$ feature channels. To ensure compatibility for concatenation, padding is applied along the time dimension, with padding sizes of 16, 32, and 48 for the three convolutional submodules. Finally, the convolutional features from all three scales are concatenated along the feature dimension, so the number of feature channels equals FC_1 and obtain $X_c \in \mathbb{R}^{FC_1\times C\times T}$. FC_1 is set to 12 in our model. By using multi-scale temporal convolutions, the model can capture features from different time windows, including short-term transient signals and long-term rhythmic fluctuations. From a frequency domain perspective, this approach enables dynamic extraction of features across different EEG frequency bands.

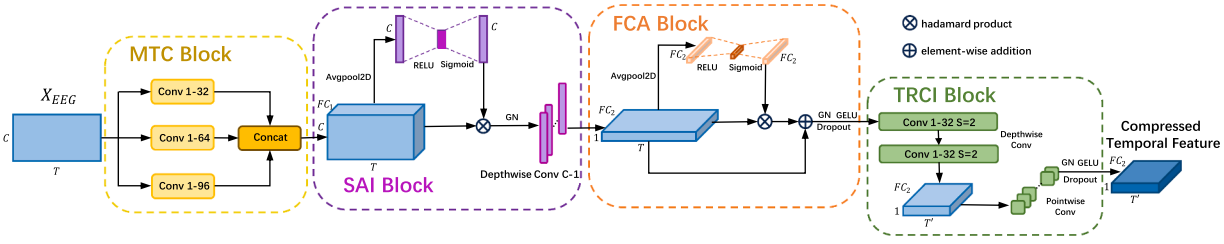

Fig. 3. The architecture of the proposed Phased Encoder. MTC: Multi-scale temporal convolution; SAI: Spatial attention and integration; FCA: Feature channel attention; TRCI: Temporal reduction and channel integration; GN: Group normalization. The Conv$x-y$ represents a convolutional layer with a kernal of size (x, y). The additional S represents the stride; if not specified, it is assumed to be 1.

2.1.2. Spatial attention and integration

This module consists of a spatial attention mechanism, specifically a Squeeze-and-Excitation (SE) module, and depthwise spatial convolution. For EEG signals, specific brain regions often show stronger responses during certain tasks, requiring higher weights for specific channels. However, these spatial weights can vary significantly across subjects and trials. To address this variability, the SE attention mechanism is introduced to adaptively adjust spatial weights (Hu, Shen, & Sun, 2018), as detailed below:

$$X_{sa} = \sigma(W_{ex}\delta(W_{sq}||_{c=1}^{C} avg(X_c))) \cdot X_c \quad (1)$$

First, average pooling (avg) is applied across the temporal and feature channel dimensions, generating a spatial feature vector of length C. This vector is then compressed by multiplying with the weight matrix W_{sq}, reducing its dimensionality to 2, and activated using the ReLU function, denoted as δ. Next, the feature vector is expanded back to C dimensions with another weight matrix W_{sq}, and the resulting weights are normalized to the range $[0, 1]$ using a Sigmoid activation, denoted as σ. These weights are then applied to the spatial dimension to emphasize the relevant regions. To address the significant inter-subject variability in EEG data, group normalization is used instead of the commonly applied batch normalization, ensuring more stable training and inference (Wu & He, 2018). Finally, spatial convolution is applied to integrate the features across all channels. The convolution kernel has a temporal size of 1 and a spatial size of C, reducing the spatial dimension to 1. Depthwise convolution is applied, with each original feature channel processed by two convolutional kernels. This results in a final output $X_s \in \mathbb{R}^{FC_2 \times 1 \times T}$, where $FC_2 = FC_1 \times 2$.

2.1.3. Feature channel attention

After integrating spatial information, the resulting multi-layer feature channels exhibit dynamic importance in classification tasks. Therefore, a Squeeze-and-Excitation (SE) attention mechanism is also introduced in the feature channel dimension, as follows:

$$X_{ca} = \sigma(W'_{ex}\delta(W'_{sq}||_{f=1}^{F_2} avg(X_s))) \cdot X_s + X_s \quad (2)$$

In contrast to the previous section, here average pooling is applied along the temporal dimension to obtain a feature channel vector of length FC_2. This vector is then compressed to one dimension using a weight matrix W'_{sq}, followed by ReLU activation. Finally, it is expanded back to FC_2 dimensions using another weight matrix W'_{ex}. The resulting feature weights are used to reweight the input along the feature dimension. Residual connections are then employed to add the original input to the weighted output, enhancing gradient flow and preserving information. This channel-level attention mechanism also models interactions across feature channels, addressing the limitation of the depthwise convolution, which lacks cross-channel interactions. Following this, group normalization is applied, and the data is passed through the GELU activation function (Hendrycks & Gimpel, 2016). Finally, dropout with a rate of 20 % is applied to reduce overfitting to individual subjects in cross-subject tasks and improve the model's generalization performance.

2.1.4. Temporal reduction and channel integration

This module further extracts temporal information and reduces dimensionality along the time axis using multi-layer convolutions, followed by pointwise convolution for feature channel integration. First, two layers of depthwise temporal convolution with a window size of 32 are applied. To reduce the computational complexity of subsequent temporal feature extraction, the convolution stride is set to 2, effectively reducing the temporal dimension. Next, pointwise convolution is employed to enable interaction and fusion across different feature channels. The output is then passed through group normalization, a GELU activation function, and a dropout layer with a rate of 20 %. This process produces the final Compressed Temporal Feature with dimensions $FC_2 \times 1 \times T'$.

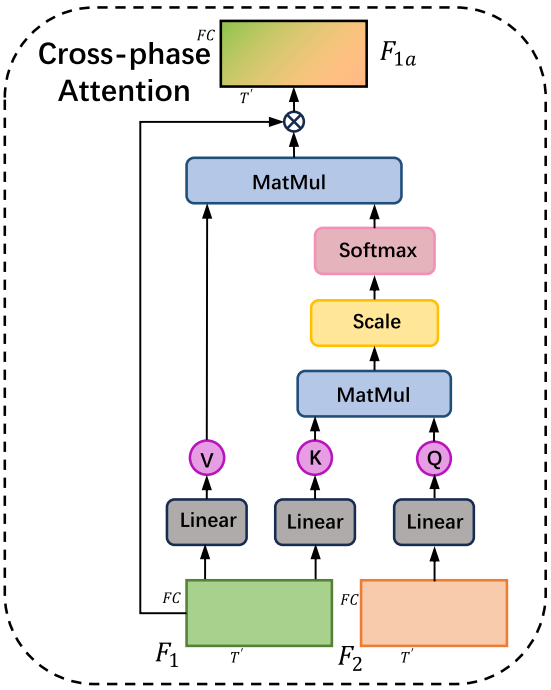

Fig. 4. The mechanism of cross-phase attention employed in our model. The features from the two phases are both processed through cross-phase attention to obtain their respective weighted features.

2.2. Temporal alignment and fusion

The temporal features of the segmented EEG signals have been extracted using the Phased Encoder described in Section 2.1. The temporal features from the two phases are first processed through a cross-attention mechanism to determine their interaction weights, which are then applied with weighting and residual connections. The features are then aligned in high-dimensional space using Maximum Mean Discrepancy (MMD) Loss to match their distributions. To preserve the temporal relationship between the two phases, the features are concatenated based on their time alignment. Finally, the concatenated features are processed through a Light temporal convolutional network (LightTCN) to extract global temporal features, followed by Binary Cross Entropy (BCE) loss to determine whether a target is detected.

2.2.1. Cross-phase attention

To achieve better fusion of the features from the two phases, the cross-phase attention mechanism is first applied to weight the features. This enables the model to enhance the feature representation of the each phase by incorporating information from the other phase, thereby improving the model's discriminative ability while processing the signal from the current phase. The cross-phase attention mechanism in this model is illustrated in Fig. 4.

Here, the features to be weighted are denoted as F_1, and the features used to extract the interaction attention are referred to as F_2. For the input F_1, two linear transformations are applied along the feature dimension to obtain the K and V. The input F_2 passes through a linear layer to generate the Q. The Q and the transpose of K are multiplied through matrix multiplication to achieve information interaction and obtain attention scores. Then, the scores are scaled by dividing by $\sqrt{F}$ and F represents the feature dimension. Softmax is then applied for normalization to generate attention weights. The V from F_1 is then multiplied by the attention weights to obtain the cross-attention weights. Finally, F_1 is weighted by the attention mechanism to produce the final

interacted feature F_{1a}. Its calculation formula is as follows.

$$F_{1a} = \left(\text{Softmax}\left(\frac{W_q(F_1) \times W_k(F_2)^T}{\sqrt{F}} \right) \times W_v(F_1) \right) \circ F_1 \tag{3}$$

$W_q(\cdot)$,$W_k(\cdot)$ and $W_v(\cdot)$ represent the linear mappings, and $\circ$ denotes the Hadamard product. Finally, the weighted features are added to the original features through a residual connection, enhancing the feature representation and preventing the loss of the original features.

2.2.2. Time-aligned concatenation

Previously, the EEG segments were divided into two phases and their independent features were extracted. However, certain EEG response features, such as attention, are global and span across the entire process. Therefore, it is essential to extract temporal features across the entire phase. To ensure the causality of the temporal sequence, the features from both phases are concatenated in a time-aligned manner, as expressed in the following equation.

$$F_{T''}(t) = \begin{cases} F_{1T'}(t), & t \in [0, \frac{1}{3}T') \\ F_{1T'}(t) + F_{2T'}\left(t - \frac{1}{3}T'\right), & t \in [\frac{1}{3}T', T') \\ F_{2T'}\left(t - \frac{1}{3}T'\right), & t \in [T', \frac{4}{3}T'] \end{cases} \tag{4}$$

The features $F_{1T'}$ and $F_{2T'}$ represent the EEG segments from the first 0.75 second and the last 0.75 second, respectively. After feature extraction using the Phased Encoder, the time dimension is reduced to T'. For time alignment, the second phase feature starts at time zero, corresponding to $T'/3$ of the first phase feature. As a result, the last two-thirds of $F_{1T'}$ overlaps with the first two-thirds of $F_{2T'}$ in time, and these sections are element-wise added together. The first one-third of $F_{1T'}$ and the last one-third of $F_{2T'}$ are then concatenated at the beginning and end, respectively. After the time-aligned concatenation, the time dimension increases from T' to $3T'/4$, denoted as T''. This time-aligned concatenation captures global temporal characteristics, enabling the model to better understand interactions between different phases and improving prediction performance.

2.2.3. Light temporal convolutional network

To extract global temporal information from the concatenated features, we use a Light temporal convolutional network (LightTCN). By leveraging causal convolutions and dilated convolutions, TCN effectively captures long-range dependencies, addressing issues like gradient vanishing or explosion commonly encountered in Recurrent neural networks (RNNs) (Bai, Kolter, & Koltun, 2018). This makes TCN particularly suitable for extracting and integrating full-stage EEG temporal features. To make the model more lightweight and reduce overfitting, we simplified the original TCN, retaining only its core components: causal dilated convolutions and 1D convolutional residual connections. The LightTCN module used in our model is shown in Fig. 5.

The input to the LightTCN is a concatenated feature matrix of size $FC \times T''$, which passes through three residual submodules. For each submodule, the input undergoes causal dilated convolutions. Causal convolutions ensure that the output at each time step depends only on the current and previous time steps, preventing information leakage from future time steps. Dilated convolutions increase the receptive field by inserting gaps in the convolution kernels, enhancing the ability to capture long-term temporal features. In this model, the dilation factor increases sequentially, with values of 1, 2, and 4 for the three submodules, respectively, expanding the temporal receptive field. After the causal dilated convolution, a ReLU activation function is applied to introduce non-linearity. Each submodule also includes a residual connection, where pointwise convolutions are applied to the input, and the output is added to the result of the previous causal dilated convolution. This allows the model to learn information at different time scales more effectively. After passing through the three residual submodules, the output is aggregated along the time dimension using average pooling,

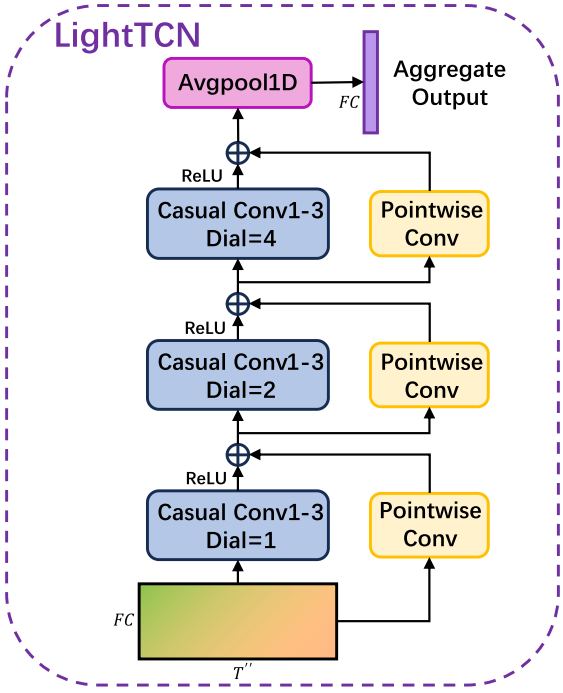

Fig. 5. The mechanism of LightTCN employed in our model.

producing a global feature vector of length FC, which is then passed through a fully connected layer for the final classification.

2.2.4. Loss function

The loss function of this model consists of two components: the Maximum Mean Discrepancy (MMD) loss for aligning the features of the two phases and the Binary Cross Entropy (BCE) loss for the final classification. Before performing the time-aligned concatenation of the two-phase features, the MMD loss is applied to align the features in the high-dimensional space (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2012). This alignment reduces the distributional differences between the two feature sets, ensuring a smoother concatenation process and enhancing the effectiveness of the resulting concatenated features. The MMD loss is calculated as follows:

$$\begin{aligned} \mathcal{L}_{\text{MMD}} = \mathbb{E}_{x,x' \sim \mathcal{P}}\left[K(x,x')\right] + \mathbb{E}_{y,y' \sim \mathcal{Q}}\left[K(y,y')\right] \\ - 2\mathbb{E}_{x \sim \mathcal{P}, y \sim \mathcal{Q}}[K(x,y)] \end{aligned} \tag{5}$$

Here, $\mathcal{P}$ and $\mathcal{Q}$ represent the distributions of the features from the two phases. x and x' denote two independent feature samples from the first phase, while y and y' denote two independent feature samples from the second phase. $\mathbb{E}[\cdot]$ denotes the expectation operator. $K(\cdot)$ is the kernel function used to map the features to a high-dimensional space; in this model, a Gaussian kernel is utilized for the mapping. The formula is as follows:

$$K(u,v) = \exp\left(-\frac{\|u - v\|^2}{2\sigma^2} \right) \tag{6}$$

σ represents the bandwidth of the Gaussian kernel, which controls the rate at which similarity decays. To achieve classification, a BCE loss function is used to measure the difference between the model's predicted probabilities and the actual labels.

$$\mathcal{L}_{\text{BCE}} = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \right] \tag{7}$$

Here, y represents the actual labels, and $\hat{y}$ represents the predicted class probabilities. The final loss function is obtained by combining the MMD loss and BCE loss with a weighted sum, as follows:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{BCE}} \tag{8}$$

In the formula, λ represents the weighting coefficient for the MMD loss. By combining the MMD loss and BCE loss with a weighted sum, the model simultaneously optimizes both phase feature alignment and classification performance during training, with λ adjusting their relative contributions.

3. Experiments

3.1. Experiment paradigm

Eight participants with normal or corrected-to-normal vision participated in the experiment. The experiments strictly followed the principles of the 2013 Declaration of Helsinki and were approved by the Research Ethics Committee of Beijing Institute of Technology with the ethical review number: BIT-EC-H-2024150. To simulate a target detection task in low-quality UAV video footage, a UAV aerial surveillance scenario over the ocean was constructed. The UAV flew at a constant speed over the sea, capturing video at a resolution of 1920×1080. The targets to be detected were aircraft and ships, concealed among clouds, islands, and waves to replicate the environmental and weather challenges faced in real-world aerial footage.

A total of 240 20-second video clips were created, with half containing targets. Each video included at most one target, which appeared randomly between 6 and 14 second into the clip, at varying locations. This randomness was designed to prevent participants from anticipating the timing and location of target appearances.

During the experiment, participants' EEG and eye movement signals were recorded. EEG data were collected using a NeuSen W64 portable wireless amplifier with 64 electrodes, following the international 10–20 system. The AFz position was used as the ground electrode, and the CPz position served as the reference electrode. The sampling frequency was set to 1,000 Hz. Eye movement data were recorded using a Tobii Pro-Fusion Screen eye tracker with a sampling rate of 60 Hz, tracking the x and y gaze coordinates on the display screen for each frame. For further details on the experimental setup and protocol, please refer to our previous work (Shi et al., 2024).

3.2. Data preprocessing

3.2.1. Initial data processing

In our experiment, baseline correction is first applied to the EEG signals, using the mean value of the 0–1 second interval as a reference to remove any offset. To reduce the impact of artifacts and noise on signal quality, we apply a bandpass filter with a 0.5-49Hz range. Next, the signal is downsampled from 1000Hz to 200Hz to reduce computational load. Next, a whole-brain common average reference is performed to minimize the influence of reference electrodes and enhance the quality of signal representation. These operations are carried out using EEGLab 2024.0 (Brunner, Delorme, & Makeig, 2013).

3.2.2. Processing with and without ICA

In non-invasive EEG experiments, some non-brain artifacts like eye movements and muscle activity often interfere with the signal, and different studies adopt various approaches to handle them. In neuroscience research, especially studies emphasizing rigor, independent component analysis (ICA) is frequently used to identify and remove artifact-related components (Lu et al., 2022b; Zhou et al., 2024). However, in real-time BCI applications, the computational efficiency of ICA for artifact removal is relatively low, and the effectiveness of artifact removal is limited, especially when the subject's training EEG segments are few, or in cross-subject experiments. Moreover, in target detection paradigms, some eye movement signals might be relevant to detecting targets, and retaining this information could potentially improve the model's performance. As a result, certain studies choose not to explicitly remove artifacts such as eye movements (Cui et al., 2023; Lu, Zhang, Chu, Liu, & Yu, 2022a).

To demonstrate the robustness of our proposed brain-inspired model, we conducted tests both with and without ICA artifact removal. This dual approach allowed us to assess the model's decoding ability for pure brain signals and its robustness in handling mixed signals containing artifacts. This demonstrates the rigor of our method in neuroscience research and its practical applicability in real-world scenarios.

For ICA processing, we applied ICA separately to each subject's EEG data to decompose the signals into independent components. Following this decomposition, we utilized the ICLabel algorithm to automatically classify these components based on their likelihood of representing neural activity or artifacts (Pion-Tonachini, Kreutz-Delgado, & Makeig, 2019). Specifically, components labeled as "EYE" (ocular artifacts) and "MUSCLE" (electromyographic artifacts) with a confidence score exceeding 90 % were identified as artifact-related components and subsequently removed. After excluding these components, we performed an inverse ICA transformation to reconstruct the EEG signals, ensuring that only neural-related activity was retained while minimizing contamination from artifacts. This process resulted in artifact-corrected EEG signals that were cleaner and more suitable for further analysis.

3.2.3. Asynchronous EEG data alignment techniques

In low-quality video target detection, an asynchronous detection issue arises due to an unpredictable delay between the target's appearance in the video and the participant's detection (Song, Yan, Tong, Shu, & Zeng, 2020). This delay complicates the accurate segmentation of ERP data. To address this challenge, we previously proposed a method for aligning Fixation-Related Potentials (FRP) with eye-tracking signals, which demonstrated its effectiveness (Shi et al., 2024) and is also applied in the current experiment.

Specifically, we compared the eye-tracking coordinates with the target's position, which was predetermined during the video creation process. When the eye-tracking coordinates fall within the target's area, the eye movement is classified as either a saccade or smooth pursuit using the velocity-threshold identification fixation classification algorithm (Prabha & Bhargavi, 2020). A saccade indicates the participant is searching for the target, while smooth pursuit signals that the target has been detected and is being tracked (Williams, 2020). The moment when the eye-tracking coordinates align with the target's position and the eye movement is classified as smooth pursuit is defined as the time of target detection. This time point is then used to segment the FRP data.

3.2.4. Temporal division of data segments

In each trial, the moment when the participant detects the target, determined based on eye movement, is set as the reference time (zero point). The EEG segment from [0,1] second is labeled as the FRP corresponding to target detection, while the [-4,-3] second segment is considered the segment for when the target was not detected. Across the 8 participants, the total number of positive and negative samples is balanced at 875 each.

3.3. Cross-subject training and evaluation

To evaluate the model's generalization ability and assess its reliability in real-world applications while increasing data diversity, a Leave-One-Out Cross-Validation (LOOCV) approach was adopted. Among the eight participants, one participant was designated as the test set in each round, while the remaining seven participants were divided into a training set (comprising six participants) and a validation set (comprising one participant), resulting in eight rounds of training and evaluation. For each training round, the model weights achieving the lowest Binary Cross Entropy loss on the validation set were selected for evaluation on the test set. Three metrics were used to evaluate the model: accuracy (Acc), true positive rate (TPR), and false positive rate (FPR). These metrics were selected to assess the overall performance of the model,

its ability to correctly classify target EEG signals, and its ability to correctly reject non-target EEG signals. The formulas for these metrics are as follows:

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$\text{TPR} = \frac{TP}{TP + FN}, \text{FPR} = \frac{FP}{FP + TN} \tag{10}$$

where TP denotes the true positive instances, TN the true negative instances, FP the false positive instances, and FN the false negative instances.

During training, the Adam optimizer was employed with a learning rate of 1e-3, and the first and second moment estimates (β_1 and β_2) were decayed with factors of 0.9 and 0.999, respectively. The batch size was set to 16. To mitigate the risk of overfitting in cross-subject training, the number of epochs per round was limited to 5. The weight coefficient of the MMD loss (λ) is set to 1.5. The model was implemented using PyTorch 1.13.1 and trained and evaluated on NVIDIA GeForce RTX 3050.

3.4. Pseudo-online test

To further evaluate the model's performance in real-world applications, we conducted a pseudo-online test. Using a sliding window of 1 second with a step size of 0.1 seconds, each small window of data was processed through the model to obtain results. To reduce false positives caused by the non-stationarity of EEG signals and enhance the robustness of the detection system, we applied a continuous hit strategy. This strategy requires that a signal segment is classified as having a target only if it is continuously detected as such above a certain threshold. In the experiment, we tested the proposed model with thresholds ranging from 3 to 7. In addition to recording accuracy, TPR, and FPR, we also compared the model's detection latency - the time difference between when the model detects a target and when the subject actually sees the target. This was used to assess the model's real-time performance in practical applications.

3.5. Evaluation under cognitive distraction

To further evaluate the model's performance under different conditions, we conducted an additional experiment with two participants performing video target detection under cognitive distraction. This aimed to assess the model's effectiveness in real-world multitasking BCI applications and explore its performance limits.

To introduce cognitive distraction, we incorporated the *n*-back paradigm as a secondary cognitive task during the primary video target detection task. Originally proposed by MIT Agelab, the *n*-back paradigm imposes cognitive load by playing a sequence of digits and requiring participants to recall previously presented digits (Mehler, Reimer, & Dusek, 2011). This method is widely used in cognitive distraction studies. In our experiment, an audio sequence of random digits (0-9) was played starting from the beginning of the video. Each digit was presented for 0.75 seconds with a 2-second interval between digits. While performing the primary low-quality video target detection task, participants were also required to recall and verbally report the digit presented *n* steps earlier. Given the high visual and cognitive load of the primary task, we used a 1-back condition to induce cognitive distraction.

For participants under cognitive distraction, 40 % of the trials were designated as distraction trials, resulting in a total of 308 positive and negative samples across both participants. Due to the extended time gap between this experiment and the previous non-distraction experiment, significant cross-session differences emerged from variations in signal acquisition equipment and environmental factors. To address this, model validation was conducted separately for each participant. Data preprocessing followed the same procedure as outlined in Sections 3.2.1 and Sections 3.2.3, with no ICA applied to simulate real-world conditions. The training and evaluation employed five-fold cross-validation, with the dataset split into training, validation, and test sets in a 3:1:1 ratio. The hyperparameter setup remained consistent with Section 3.3, but given the limited dataset, the number of epochs was set to 30.

3.6. Comparison with baseline models

In this study, we evaluated the proposed model against several baseline models, both with and without ICA artifact removal. The baselines included the classical HDCA model from the RSVP paradigm (Gerson et al., 2006); temporal networks such as GRU (Cho et al., 2014) and LSTM (Wang, Jiang, Liu, Shang, & Zhang, 2018); convolution-based networks like DeepConvNet (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018), and EEG-Inception (Santamaria-Vazquez, Martinez-Cagigal, Vaquerizo-Villar, & Hornero, 2020); as well as hybrid deep learning models with structures similar to the proposed model, though employing different fusion strategies - such as the graph-convolutional STGCN (Yu, Yin, & Zhu, 2017) and the time-frequency fusion transformer TFF-Former (Li et al., 2022b).

4. Results and discussion

4.1. Phased neural signature results

For video target detection tasks, the brain's response is multi-phased, which served as inspiration for the development of the proposed model. Neuroscientific research highlights the involvement of three key brain regions in video target detection: the posterior parietal cortex, parts of the higher visual cortex (including areas V2, V3, and MT), and the prefrontal cortex. The posterior parietal cortex primarily generates the P300 waveform as a surprise response when identifying and detecting targets, which occurs predominantly in the early stages of target detection (Pogarell et al., 2011). The V2, V3, and MT areas are mainly involved in spatial target perception and tracking. The V2 area focuses on primary visual feature detection and disparity processing, the V3 area handles initial perception of dynamic objects (Kaas, 2003; Polich, 2007), and the MT area processes advanced motion perception, including speed, direction, and visual tracking (Born & Bradley, 2005; Jahn et al., 2012). These responses become more prominent during the later stages of target tracking. The prefrontal cortex, responsible for attention regulation, exhibits significant responses throughout all phases of target detection (Knight, 1994; Martinez-Trujillo, 2022).

To analyze the phased activity of these three brain regions using EEG signals, representative channels corresponding to each region were selected. Specifically, Pz represents the posterior parietal cortex, O2 represents the spatial visual cortex, and Fpz represents the prefrontal cortex. To reduce interference from artifacts such as eye movements in EEG representations, we used independent component analysis (ICA) to remove artifact-related components. Both time-domain and frequency-domain analyses were conducted: the time-domain analysis utilized Fixation-Related Potentials (FRP), while the frequency-domain analysis employed Fixation-Related Spectral Perturbation (FRSP).

4.1.1. FRP Results

The FRP analysis was conducted by averaging all trials across the eight participants. Fig. 6 shows the FRP of three representative channels: Pz, O2, and Fpz.

For the Pz channel, a prominent positive waveform is observed in the 0-0.75 second interval, with a peak amplitude of approximately 3 μV, reflecting a significant P300 waveform. This indicates that the brain response during this phase is primarily driven by recognition and the surprise associated with target appearance(Pogarell et al., 2011). Notably, due to the use of eye-tracking for FRP alignment, the P300 response appears slightly earlier than usual.

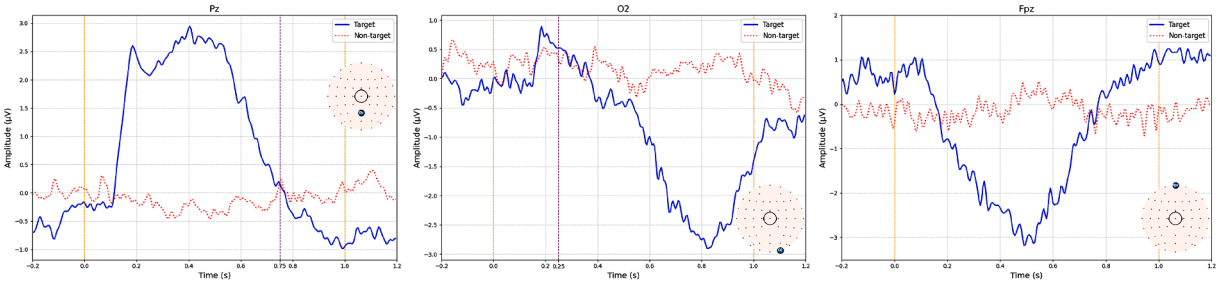

Fig. 6. Time-domain responses provide the basis for EEG phase segmentation. FRP of Pz, O2, and Fpz channels. The blue line represents the EEG response when the participant detects the target, with the 0–1 second interval marked by yellow dashed lines (0 second indicating target appearance, and the 0–1 second window representing the main FRP period). The red dashed line represents the EEG response when no target is detected for comparison. The main P300 response at Pz occurs between 0-0.75 second, with the purple dashed line marking 0.75 second. The main negative wave response at O2 appears between 0.25-1 second, with the purple dashed line at 0.25 second.

For the O2 channel, a strong negative waveform is observed between 0.25-1 second, with a peak amplitude of approximately 2.9 μV. This negative waveform gradually recovers after 0.8 second. As O2 is located in the secondary visual cortex and dorsal stream regions associated with spatial attention, this response is likely driven by spatial tracking of the target after detection (Born & Bradley, 2005). In contrast, the symmetrical O1 channel on the left side does not show a significant negative waveform, which may be due to the brain's right-hemisphere dominance in spatial attention tasks (Hellige, 1996). This asymmetry further supports the conclusion that the negative waveform at O2 is primarily related to visual spatial attention.

For the Fpz channel, a significant negative waveform is observed throughout the 0–1 second interval, with a maximum drop of approximately 4.2 μV, showing a trend of initial decline followed by recovery. As Fpz is located in the prefrontal cortex, this response is likely associated with cognitive processes such as attention maintenance, target confirmation, and subsequent decision-making (Knight, 1994; Martinez-Trujillo, 2022).

The FRP analysis indicates that the brain's response after target detection can be divided into three distinct phases: the recognition phase (0-0.75 second) represented by Pz, the spatial tracking phase (0.25-1 second) represented by O2, and the attention phase spanning the entire period represented by Fpz. Inspired by these findings, we developed a phase-based EEG classification model for video target detection.

4.1.2. FRSP Results

To further validate the phase-based cognitive model of the brain in video target detection, we performed a frequency-domain analysis using FRSP for the three corresponding brain regions (Pascual-Marqui et al., 2002), as shown in Fig. 7. Using the 4 seconds before fixation as the baseline, we applied a 3-cycle wavelet transform with a scaling factor of 0.5, a 200-time-point temporal window, and a frequency range of 0.5–13 Hz. This range covers the delta (0.5-4 Hz), theta (4–8 Hz), and alpha (8–13 Hz) bands.

For the Pz channel, a strong response is observed in the 3–13 Hz range within 0.4 second of target detection, primarily in the theta and alpha bands. These bands are associated with top-down cognitive processing in the brain (Min & Park, 2010), which, in this task, corresponds to target recognition. Additionally, a prolonged delta-band response is present in the 0-0.75 second interval, potentially reflecting the synchronization and integration of multisensory inputs across brain regions (Hermer-Vazquez, Hermer-Vazquez, & Srinivasan, 2009).

For the O2 channel, changes in band power are primarily concentrated in the 0.5–8 Hz range, corresponding to the delta and theta bands. After target detection, particularly starting at 0.25 second, the power in the delta to theta bands gradually increases and remains at a high level. Between 0.8 and 1 second, the range of high-power frequency bands reaches its peak, with a baseline ratio of approximately 4 dB. The low power observed during the early phase may reflect low-level visual feature extraction associated with the primary visual cortex (V1). The increased delta and theta band power in the mid-to-late phase likely corresponds to shifts and concentration of spatial attention (Jiang, Zhang, & Yu, 2021), as well as the dynamic integration and spatial tracking of the video target (Senoussi, Moreland, Busch, & Dugué, 2019).

For the Fpz channel, strong responses are observed throughout the 1-second interval. From 0 to 0.4 second, significant power appears in the 5–11 Hz range (theta and low-frequency alpha bands). This likely reflects attention shifts triggered by target appearance, interactions between the prefrontal cortex and hippocampus (Roy, Svensson, Mazeh, & Kocsis, 2017), and attentional resource allocation involving regions such as the parietal cortex (Sauseng et al., 2005). From 0.4 to 1 second, higher power is observed in the 0.5-4 Hz and 11–13 Hz ranges (delta and high-frequency alpha bands). The delta response may indicate long-range coordination of the prefrontal cortex with other brain regions (Hermer-Vazquez et al., 2009) and could relate to emotional reactions, such as task completion satisfaction (Cavanagh, 2015). The high-frequency al-

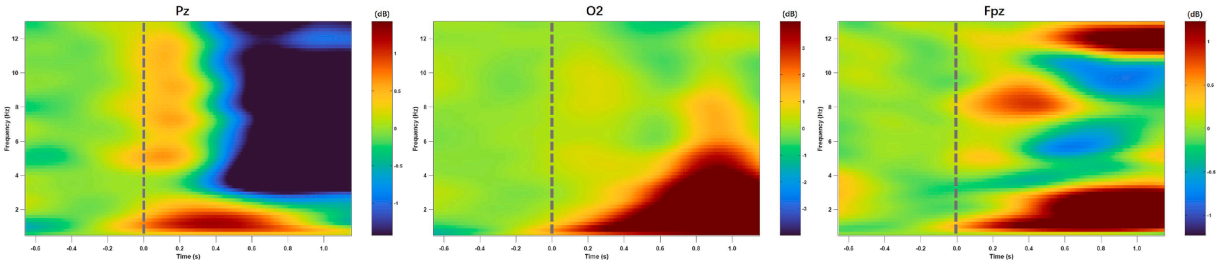

Fig. 7. Frequency-domain responses provide the basis for EEG phase segmentation. FRSP of the Pz, O2, and Fpz channels. The 0-second mark, indicating the moment the participant detected the target, is marked with a dashed line. Red and blue areas in the figure represent power levels higher and lower than the baseline, respectively.

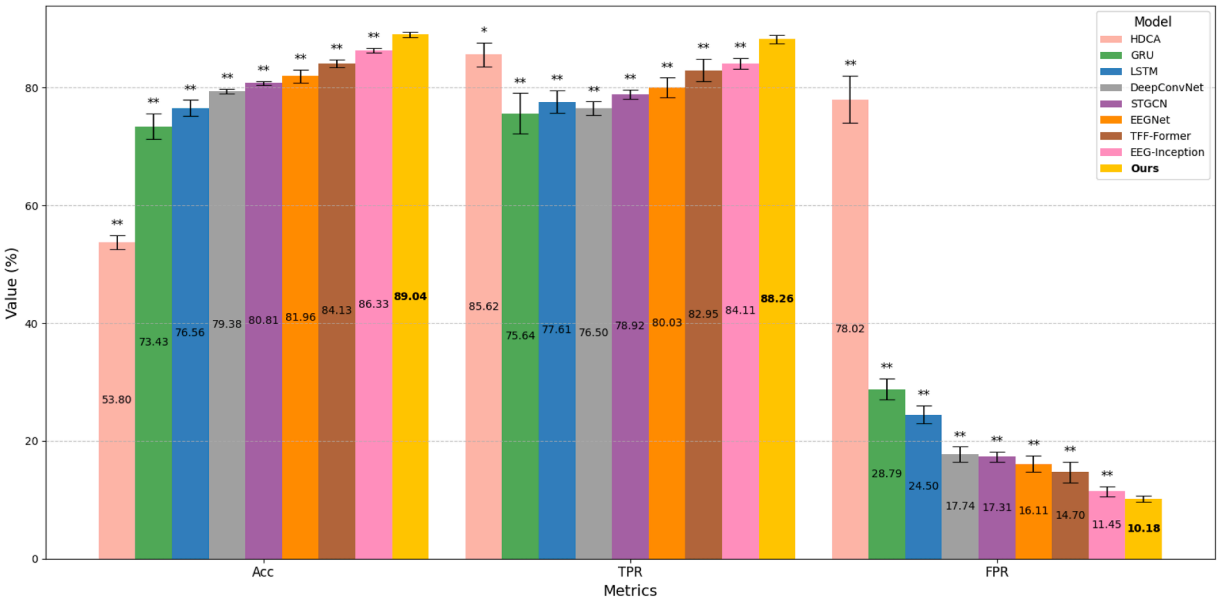

Fig. 8. Comparison of accuracy, true positive rate, and false positive rate for different models using data with ICA. Error bars show the 95 % confidence interval based on the t-distribution over ten runs. * indicates $p < .05$, and ** indicates $p < .01$ in Wilcoxon signed-rank test.

pha response, associated with sustained attention (Clayton, Yeung, & Kadosh, 2015), aligns with spatial tracking activities observed in the O2 channel and may also support target-background suppression.

From the above FRSP analysis, in the frequency domain, Pz demonstrates a primary response between 0-0.75 second, while O2 exhibits an increasing response starting from 0.25 second and sustaining thereafter. Although Fpz responses differ in frequency bands between the early and later phases, its overall response spans the entire 0–1 second interval. These findings offer additional neuroscientific support for the design of a phase-based EEG classification network.

4.2. Test results

4.2.1. Results with ICA

Fig. 8 presents a visual comparison of accuracy, true positive rate (TPR), and false positive rate (FPR) between our proposed model and baseline models under ICA processing. To ensure reliability, the results represent the average of ten independent runs. Our model demonstrates superior performance across all metrics, achieving an accuracy of 89.04 %, TPR of 88.26 %, and FPR of 10.18 %. Compared to the second-best performing model, EEG-Inception, our model shows improvements of 2.71 %, 4.15 %, and 1.27 % in accuracy, TPR, and FPR, respectively. To validate the statistical significance of our results, we conducted Wilcoxon signed-rank tests on the ten-run results for each model. The tests revealed that our model's performance metrics were significantly different ($p < .01$) from those of other models, except for the TPR comparison with HDCA, which showed significant but less pronounced difference ($p < .05$). Notably, HDCA exhibited a substantially higher false alarm rate of 78.02 %. Overall, our network achieves an optimal balance between TPR and FPR.

To demonstrate the lightweight design of our model, we compare its parameter count with that of other baseline models, as shown in Fig. 9. Our model has only 20,156 parameters, second only to EEGNet, but significantly outperforms it in performance. The second-best model, TFF-Former, has a massive parameter count exceeding 1.5 million, further highlighting the advantage of our model in achieving high accuracy while maintaining a lightweight design.

4.2.2. Results without ICA

Compared to Figs. 8, A.13 shows the classification results for each model using EEG data without ICA artifact removal. Our proposed model still achieves the best performance across accuracy, TPR, and FPR, with results of 96.32 %, 95.65 %, and 3.01 %, respectively, surpassing the second-best model, EEG-Inception, by 2.25 %, 2.16 %, and 2.34 %. After running each model ten times and performing the Wilcoxon signed-rank test, our model shows a highly significant difference compared to all other models ($p < .01$). The error bars indicate that our model exhibits the least variability, highlighting its robustness.

Notably, without ICA artifact removal, our model outperforms the results obtained with ICA by 7.41 % in accuracy. To further demonstrate that potential artifacts, such as eye movement signals, can enhance the model's performance, we computed the difference between EEG signals with and without ICA processing to isolate the artifact signals. By inputting these artifacts into our model, the cross-subject accuracy reached 77.44 %, indicating that the artifacts indeed contribute to classification. We hypothesize that this improvement may be due to a shift in the participant's eye movement pattern from saccadic searching to fixation upon detecting the target, which serves as additional evidence of target detection. Therefore, in real-world target detection applications, it may be beneficial to consider not performing ICA processing, as potential artifacts could enhance model performance.

4.2.3. Pseudo-online test results

Fig. 10 presents the pseudo-online test results, showing accuracy, TPR, FPR, and detection latency at consecutive hit thresholds ranging

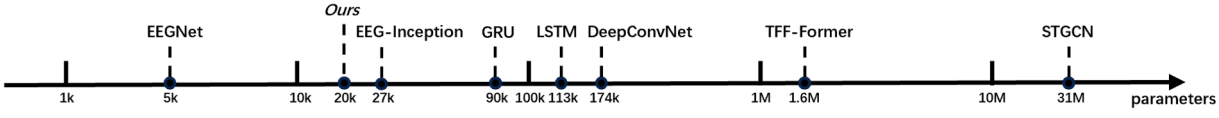

Fig. 9. Parameters comparison of different models.

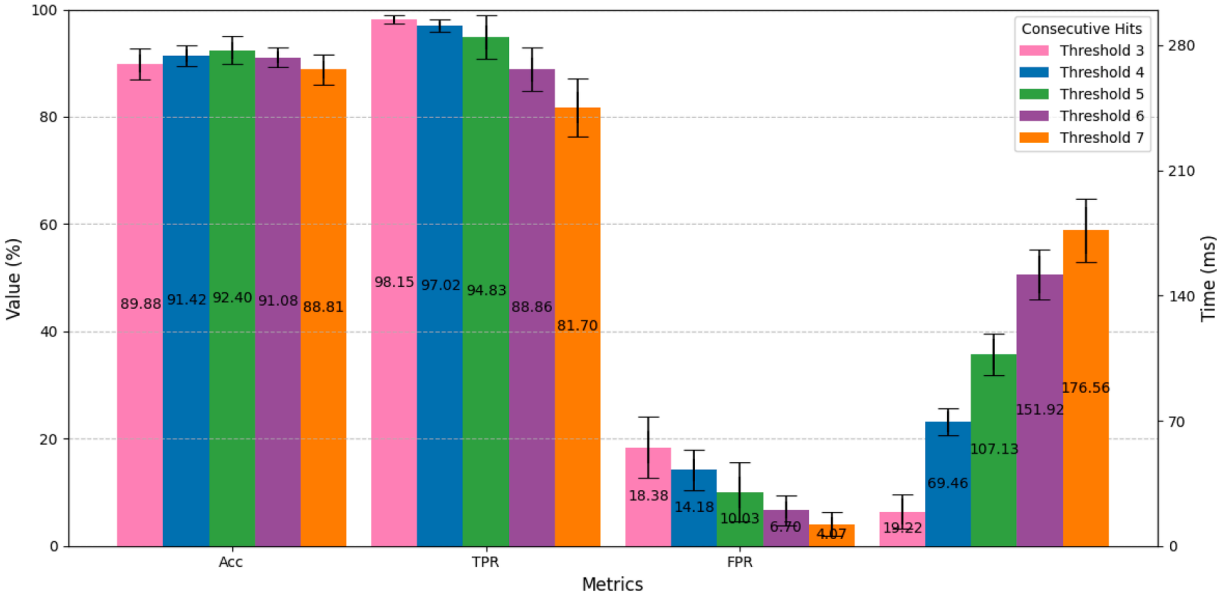

Fig. 10. The proposed model's performance in the pseudo-online test based on different consecutive hit strategies, showing the variations in accuracy, hit rate, false positive rate, and detection delay. Error bars show the 95 % confidence interval based on the t-distribution across different subjects.

from 3 to 7. As the threshold increases, which implies a higher requirement to confirm a target, the FPR decreases, but the TPR also decreases and the detection latency rises. Thus, an excessively high threshold may compromise system sensitivity and real-time performance. The figure indicates that thresholds of 4, 5, and 6 yield high accuracy (over 90 %) and a good balance between TPR and FPR. In practical applications, a threshold of 4 is recommended for high sensitivity and low latency, while a threshold of 6 is preferable for lower false positives; a threshold of 5 provides a balanced trade-off.

4.2.4. Results under cognitive distraction

Fig. A.14 presents the model accuracy results for two additional participants under cognitive distraction. While the overall model performance significantly declined due to the distraction, our model still outperformed all others by a substantial margin. It achieved an average accuracy of 75.50 % and 68.25 % for the two participants, exceeding the second-best model by 18.83 % and 9.53 %, respectively. Most other models struggled, with accuracy hovering around 50 %, indicating that many models failed to classify targets effectively under cognitive distraction. This highlights the considerable challenge introduced by cognitive distraction in video target detection decoding. To assess statistical significance, a Wilcoxon signed-rank test was conducted over ten runs, confirming that our model demonstrated a significant difference ($p < .05$) compared to all other models. These results validate the robustness of our model in handling cognitive distraction, making it well-suited for real-world multitasking scenarios.

4.3. Ablation study

4.3.1. Ablation on phase lengths

To investigate the impact of different phase time divisions on model performance, we conducted an ablation study by varying the lengths of the pre-phase and post-phase while keeping the total time duration fixed at 1 second. Specifically, we tested various phase lengths from 0.5 s to 0.8 s with a step size of 0.05 s, measured after 0 s and before 1 s, respectively. To highlight the model's practical applicability, all ablation experiments in this paper were conducted using data without ICA artifact removal. The experimental results are shown in Table 1.

The ablation study indicates that when the pre-phase and post-phase are each set to 0.75 second, the model achieves the best performance across all metrics, including accuracy, TPR, and FPR. This phase division aligns well with the temporal characteristics of the neural representations discussed in Section 4.1, further reinforcing the interpretability of the phase-based model in relation to brain mechanisms.

4.3.2. Ablation on phased encoder

To evaluate the effectiveness of each block in the proposed Phased Encoder, ablation experiments were conducted for each component. For the MTC block, the effect of multi-scale convolution was compared by replacing the three combined scales (kernels of sizes 32, 64, and 96) with individual convolutions of each scale separately while keeping the total number of kernels constant. For the SAI block and FCA block, the spatial attention and channel attention mechanisms were removed, respectively. For the TRCI block, the pointwise convolution for feature channel interaction was excluded. The results of these ablation experiments, compared to the full model, are summarized in Table 2.

For the MTC block, the use of multi-scale convolution improved accuracy by 1.44 %, 1.78 %, and 1.3 % compared to using only the 32, 64, and 96 kernel convolutions, respectively. Notably, compared to the 96-kernel convolution, multi-scale convolution maintained high accuracy while reducing the number of parameters, demonstrating its effectiveness. For the SAI and FCA blocks, spatial attention and channel attention improved accuracy by 1.92 % and 2.62 %, respectively, showing that leveraging attention mechanisms to focus on key brain regions and features relevant to target classification is an effective strategy. For the TRCI block, the pointwise convolution improved accuracy by 1.18 %, highlighting the importance of cross-channel feature interactions. Overall, the four blocks of the Phased Encoder contributed improvements

Table 1
Results of ablation on different phase lengths.

Phase Lengths (s)	Acc (%)	TPR (%)	FPR (%)
0.5	95.08 ± 0.10	94.19 ± 0.17	4.02 ± 0.25
0.55	94.31 ± 0.19	93.63 ± 0.60	5.02 ± 0.24
0.6	94.75 ± 0.49	92.84 ± 0.85	3.35 ± 0.13
0.65	95.28 ± 0.25	94.71 ± 0.26	4.14 ± 0.69
0.7	95.73 ± 0.33	95.61 ± 0.58	4.15 ± 0.26
0.75	**96.32 ± 0.13**	**95.65 ± 0.29**	**3.01 ± 0.26**
0.8	94.71 ± 0.32	93.66 ± 0.13	4.24 ± 0.55

Table 2
Results of ablation on Phased Encoder.

Block	Configuration	Acc (%)	TPR (%)	FPR (%)
MTC	32 filters only	94.88 ± 0.13	95.68 ± 0.32	5.93 ± 0.52
	64 filters only	94.54 ± 0.16	92.91 ± 0.48	3.83 ± 0.17
	96 filters only	95.02 ± 0.25	94.34 ± 0.48	4.31 ± 0.14
SAI	Without spatial attention	94.40 ± 0.22	93.91 ± 0.21	5.12 ± 0.24
FCA	Without channel attention	94.70 ± 0.42	94.31 ± 0.44	4.91 ± 0.39
TRCI	Without pointwise convolution	95.14 ± 0.29	93.78 ± 0.65	3.51 ± 0.67
	Full	**96.32 ± 0.13**	**95.65 ± 0.29**	**3.01 ± 0.26**
	EEGNet	93.23 ± 0.12	91.70 ± 0.29	5.23 ± 0.30

across accuracy, TPR, and FPR, validating the importance of each component in the model.

To validate the necessity of the proposed Phase Encoder, we replaced it with the lightweight and widely used EEGNet as the phase feature extractor. The results show that our Phase Encoder signally outperforms EEGNet across all three metrics, with accuracy improving by 3.09 %, demonstrating its superiority.

4.3.3. Ablation on feature alignment and fusion

In the two-phase feature alignment process, we compared the effects of different loss functions on alignment. In addition to the previously mentioned MMD loss, we experimented with cosine similarity loss (Salton, Wong, & Yang, 1975), KL divergence loss (Kullback & Leibler, 1951), DTW (Dynamic Time Warping) loss (Müller, 2007), and OT (Optimal Transport) loss (using the Wasserstein distance) (Peyré, Cuturi et al., 2019). Among these five losses, cosine similarity is commonly used for aligning feature directions, KL divergence for aligning probability distributions, DTW loss for finding the optimal alignment path via dynamic programming, and OT loss for computing the optimal match between distributions based on optimal transport theory. MMD loss is typically used for aligning different high-dimensional distributions. To balance the alignment loss with the final binary cross-entropy classification loss during training, we applied a weighting factor λ to the alignment loss. We varied λ from 0.25 to 2 in increments of 0.25 and recorded the impact of each loss on model accuracy, as shown in Fig. 11.

MMD, OT, and DTW losses all yielded good alignment results, with similar accuracy levels, making them all viable in practice. However, at a weight of 1.5, MMD loss achieved a peak accuracy compared to the other two, likely because, unlike OT loss, MMD does not overly focus on matching individual sample points but instead emphasizes global features. Moreover, compared to DTW loss, MMD loss minimizes excessive signal distortion, preserving the physiological consistency of the signal. Additionally, the high computational complexity of DTW loss poses challenges for model training. Cosine similarity loss performed slightly worse, with accuracy around 94 %, possibly because EEG signals exhibit highly nonlinear variations and complex temporal dependencies, which cosine similarity-calculating only the angle between vectors-fails to capture effectively. Finally, KL divergence loss produced the poorest results, with accuracy fluctuating around 90 %, likely due to the inherent asymmetry between the features of the two phases, rendering KL divergence less applicable.

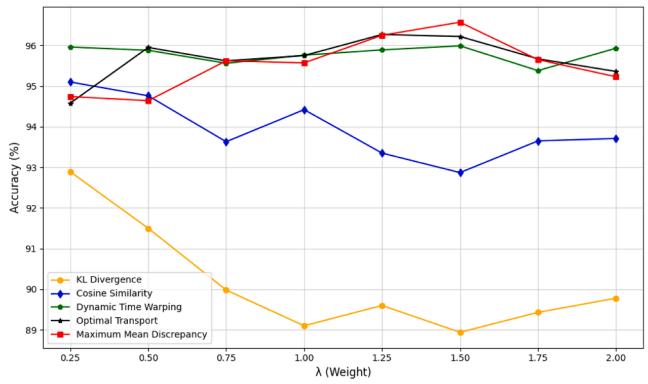

Fig. 11. The impact of different alignment loss functions and their weights in total loss function.

To validate the effectiveness of each component in the feature alignment and fusion module, we conducted ablation experiments on the cross-phase attention, time-aligned attention, and LightTCN modules. For the cross-phase attention, the module was removed entirely. For time-aligned concatenation, two alternatives were tested: direct addition and direct concatenation of features without temporal alignment. For LightTCN, we replaced it with commonly used temporal information extraction networks, GRU (Cho et al., 2014) and LSTM (Wang et al., 2018), with the same number of hidden layers (24) as LightTCN. The results are summarized in Table 3.

The cross-phase attention module improved accuracy by 1.38 %, highlighting the importance of dynamically adjusting feature weights across phases. Time-aligned attention achieved 0.77 % and 1.18 % higher accuracy compared to direct addition and direct concatenation, respectively, demonstrating that temporal alignment effectively utilizes global temporal features and avoids semantic misalignment. To further explore the influence of temporal dependency between the pre- and post-phases on the model, we conducted an ablation experiment by reversing the fusion order during alignment. The results showed a modest 0.63 % drop in accuracy, indicating that the temporal sequence does contribute some dependency cues. However, this slight decline also underscores that our phase-based partitioning remains the primary driver of performance, as the pre- and post-phase features are effectively aligned in a shared high-dimensional space. LightTCN outperformed GRU and LSTM by 3.09 % and 1.51 %, respectively, likely due to the causal dilated convolutions in LightTCN, which effectively extract both local and global temporal features. Additionally, its low parameter count helps reduce the risk of overfitting.

4.4. Visualization of feature alignment and fusion

To further demonstrate the effectiveness of feature alignment and fusion across the two phases, we visualized the features of an unseen participant (excluded from model training) using t-SNE. The visualizations include features before alignment, after alignment, and after fusion, with each point representing the projection of a single data segment, as shown in Fig. 12.

In Fig. 12 (a), the features from the two phases form separate clusters, with initial separation between the two classes within each cluster. In Fig. 12 (b), after alignment, the feature distributions of the two phases become more similar, and the two clusters are closer to each other, val-

Table 3
Results of ablation on feature alignment and fusion.

Block	Configuration	Acc (%)	TPR (%)	FPR (%)
Cross-phase attention	Without attention	94.94 ± 0.32	95.11 ± 0.30	5.22 ± 0.66
Time-aligned concatenation	Direct addition	95.55 ± 0.29	95.22 ± 0.53	4.11 ± 0.16
	Direct concatenation	95.14 ± 0.30	94.14 ± 0.17	3.86 ± 0.74
	Reversed alignment	95.69 ± 0.27	95.16 ± 0.06	3.78 ± 0.57
LightTCN	GRU	93.23 ± 0.21	93.79 ± 0.35	7.34 ± 0.75
	LSTM	94.81 ± 0.10	94.19 ± 0.36	4.56 ± 0.27
	Full	**96.32 ± 0.13**	**95.65 ± 0.29**	**3.01 ± 0.26**

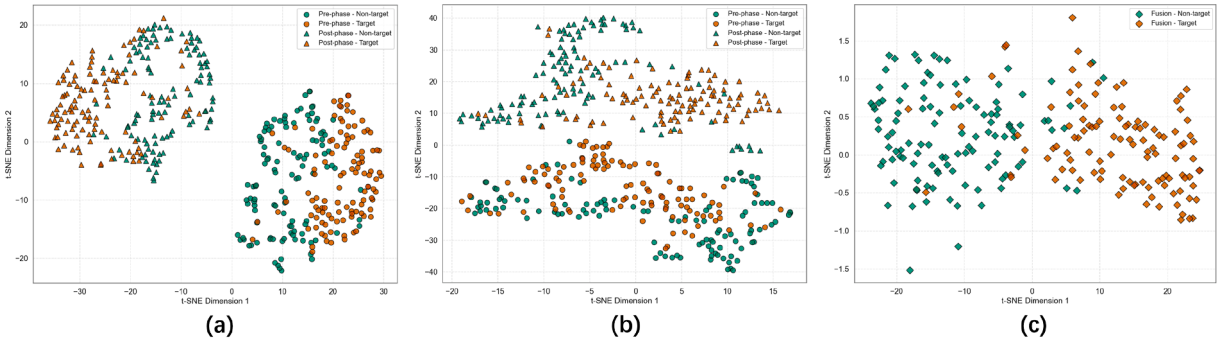

Fig. 12. (a) t-SNE of pre-phase and post-phase features before alignment; (b) t-SNE of pre-phase and post-phase features after alignment; (c) t-SNE of features after alignment and fusion.

idating the effectiveness of the alignment process. In Fig. 12 (c), after cross-phase attention and temporal fusion with LightTCN, the combined features show greater separability, demonstrating the effectiveness of the feature fusion process.

4.5. Limitations and real-world considerations

While the proposed model demonstrates strong robustness and generalizability across ICA-processed, raw, and cognitively distracted conditions, several limitations remain to be addressed in future work. First, although ICA processing helps remove common artifacts such as eye blinks and muscle activity, it cannot eliminate all sources of noise. Real-world signals often contain complex and unpredictable artifacts, including those from motion, sweating, or electrical interference, which may degrade performance in practical deployment. Second, inter-session variability, referring to the changes in brain signals recorded from the same subject across different sessions or days, has not been evaluated. This variability is known to pose significant challenges for model generalization and should be addressed in future evaluations. Third, all experiments were conducted using a single EEG recording device with a fixed number of channels. The model's compatibility with other EEG systems that differ in hardware specifications, such as channel count, sampling rate, and electrode positions, remains to be validated. Future work will consider cross-device testing and adaptation to improve deployment flexibility and robustness.

5. Conclusion

In this paper, we propose a novel brain-inspired phased temporal encoding and alignment fusion algorithm for EEG-based classification in low-quality video target detection. By analyzing the FRP and FRSP of EEG signals in both time and frequency domains, we demonstrate that the brain's response to target detection in low-quality videos can be roughly divided into three phases: early target recognition, later target spatial tracking, and global attention focus across the entire duration. Inspired by this brain mechanism, we introduce a model that splits EEG signals into early and later phases, which are processed separately by the proposed encoder. Using multi-scale convolution, spatial attention, channel attention, and feature integration, phase-specific features are extracted. These phase features are then integrated through cross-phase attention, MMD loss, time-aligned concatenation, and LightTCN for global temporal feature extraction, resulting in full-phase features for classification.

To comprehensively evaluate the superiority and robustness of the proposed model, we conducted cross-subject experiments under multiple settings, including ICA-processed and raw data, representing clean and real-world brain signals. The model achieved accuracies of 89.04 % with ICA and 96.32 % without ICA, both significantly outperforming baseline methods. Under cognitive distraction, the model maintained strong performance with accuracies of 75.50 % and 68.25 % on two independent subjects. In pseudo-online experiments, it achieved 92.40 % accuracy, further confirming its real-world potential. Ablation studies confirmed the contributions of the phase-based model, phased encoder, and phase feature alignment and fusion. Feature visualization using t-SNE further illustrated the clear discriminative power of the aligned features.

Beyond statistical improvements, the model offers practical advantages essential for real-world brain-computer interface applications. Its robustness across preprocessing pipelines, resilience under distraction, and reliable pseudo-online performance reflect strong generalizability. The lightweight architecture with fewer parameters also enables efficient deployment on portable or embedded systems. Together, these strengths highlight the model's potential to advance the development of more reliable and user-friendly neural interfaces.

This work is the first to combine the brain mechanisms of low-quality target detection with EEG-based detection algorithms, offering novel insights for the application of BCI technology in complex scenarios. By simulating the brain's phased information processing mechanism, we propose a brain-inspired algorithm that integrates brain science with ar-

tificial intelligence and other fields, advancing AI towards more human-centric and biologically informed approaches.

CRediT authorship contribution statement

Dehao Wang: Conceptualization, Methodology, Software, Writing – original draft; **Jianting Shi:** Data curation, Software; **Manyu Liu:** Data curation, Validation; **Wenao Han:** Formal analysis, Visualization; **Luzheng Bi:** Funding acquisition, Writing – review & editing; **Weijie Fei:** Project administration, Writing – review & editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported in part by the Basic Research Plan under Grant JCKY2022602C024.

Appendix A. Results of different models under various experiments

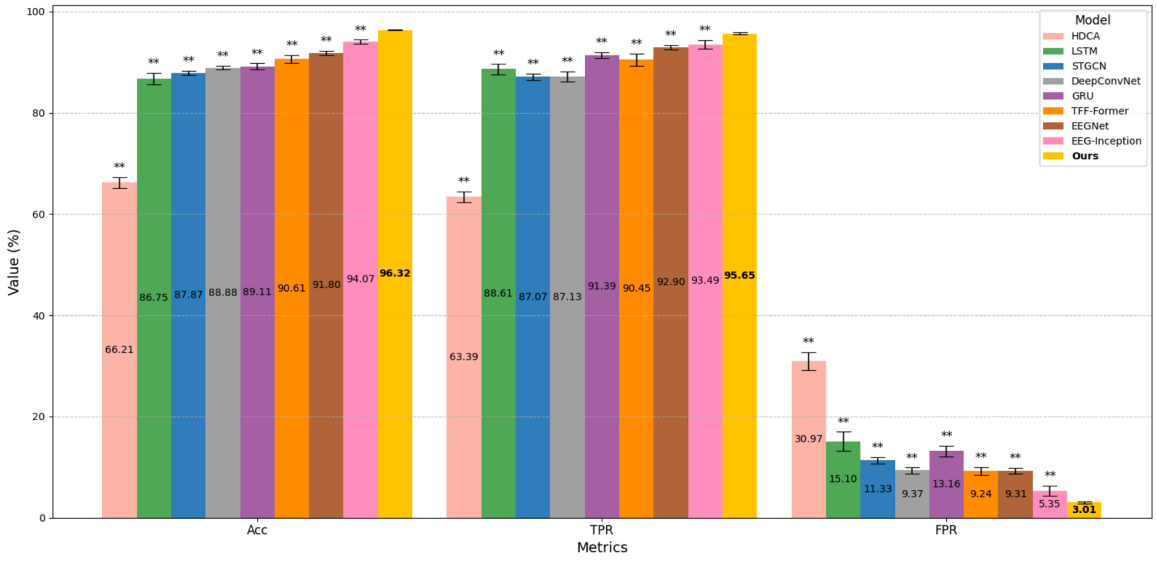

Fig. A.13. Comparison of accuracy, true positive rate, and false positive rate for different models using data without ICA. Error bars show the 95 % confidence interval based on the t-distribution over ten runs. ** indicates $p < .01$ in Wilcoxon signed-rank test.

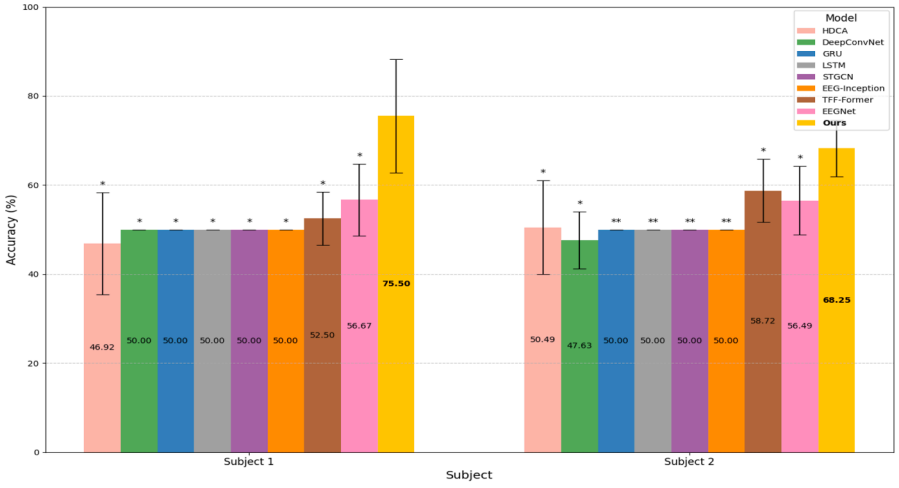

Fig. A.14. Comparison of accuracy of different models for two cognitively distracted subjects. Error bars show the 95 % confidence interval based on the t-distribution over ten runs. * indicates $p < .05$, and ** indicates $p < .01$ in Wilcoxon signed-rank test.

References

Alawad, W., Halima, N. B., & Aziz, L. (2023). An unmanned aerial vehicle (UAV) system for disaster and crisis management in smart cities. *Electronics*, *12*(4), 1051.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289.

Barachant, A., & Congedo, M. (2014). A plug&play p300 BCI using information geometry. arXiv preprint arXiv:1409.0107.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. *Annual Review of Neuroscience*, *28*(1), 157–189.

Brunner, C., Delorme, A., & Makeig, S. (2013). Eeglab–an open source matlab toolbox for electrophysiological research. *Biomedical Engineering/Biomedizinische Technik*, *58*(SI-1-Track-G), 000010151520134182.

Cavanagh, J. F. (2015). Cortical delta activity reflects reward prediction error and related behavioral adjustments, but at different times. *NeuroImage*, *110*, 205–216.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Choi, S. H. (2024). Spiking neural networks for biomedical signal analysis. *Biomedical Engineering Letters*, *14*(5), 955–966.

Clayton, M. S., Yeung, N., & Kadosh, R. C. (2015). The roles of cortical oscillations in sustained attention. *Trends in Cognitive Sciences*, *19*(4), 188–195.

Cui, Y., Xie, S., Xie, X., Zheng, D., Tang, H., Duan, K., Chen, X., & Jiang, Y. (2023). Lder: A classification framework based on erp enhancement in rsvp task. *Journal of Neural Engineering*, *20*(3), 036029.

Fang, W., Zhang, G., Zheng, Y., & Chen, Y. (2023). Multi-task learning for UAV aerial object detection in foggy weather condition. *Remote Sensing*, *15*(18), 4617.

Feng, D., Harakeh, A., Waslander, S. L., & Dietmayer, K. (2021). A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 9961–9980.

Garvanov, I., Garvanova, M., Ivanov, V., Chikurtev, D., & Chikurteva, A. (2024). Drone detection based on image processing. In *2024 23rd international symposium on electrical apparatus and technologies (SIELA)* (pp. 1–5). IEEE.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, *31*.

Gerson, A. D., Parra, L. C., & Sajda, P. (2006). Cortically coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *14*(2), 174–179.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, *13*(1), 723–773.

Hellige, J. (1996). Hemispheric asymmetry for visual information processing. *Acta Neurobiologiae Experimentalis*, *56*(1), 485–497.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.

Hermer-Vazquez, R., Hermer-Vazquez, L., & Srinivasan, S. (2009). A putatively novel form of spontaneous coordination in neural activity. *Brain Research Bulletin*, *79*(1), 6–14.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Jahn, G., Wendt, J., Lotze, M., Papenmeier, F., & Huff, M. (2012). Brain activation during spatial updating and attentive tracking of moving targets. *Brain and Cognition*, *78*(2), 105–113.

Jiang, Y., Zhang, H., & Yu, S. (2021). Changes in delta and theta oscillations in the brain indicate dynamic switching of attention between internal and external processing. In *4th International conference on biometric engineering and applications* (pp. 25–31).

Kaas, J. H. (2003). Early visual areas: V1, v2, v3, DM, DL, and MT. In *The primate visual system* (pp. 138–158). CRC Press.

Knight, R. T. (1994). Attention regulation and human prefrontal cortex. In *Motor and cognitive functions of the prefrontal cortex* (pp. 160–173). Springer.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

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, D., Zhang, Z., Wang, B., Yang, C., & Deng, L. (2022a). Detection method of timber defects based on target detection algorithm. *Measurement*, *203*, 111937.

Li, X., Wei, W., Qiu, S., & He, H. (2022b). Tff-former: Temporal-frequency fusion transformer for zero-training decoding of two bci tasks. In *Proceedings of the 30th ACM International conference on multimedia* (pp. 51–59).

Liu, S., Wang, S., Shi, W., Liu, H., Li, Z., & Mao, T. (2019). Vehicle tracking by detection in UAV aerial video. *Science China. Information Sciences*, *62*(2), 24101.

Lu, G., Zhang, Y., Chu, X., Liu, Y., & Yu, Y. (2022a). Combining multi-scale convolutional neural network and transformers for EEG-based RSVP detection. In *2022 37th youth academic annual conference of chinese association of automation (YAC)* (pp. 426–431). IEEE.

Lu, R., Zeng, Y., Zhang, R., Yan, B., & Tong, L. (2022b). Sast-gcn: Segmentation adaptive spatial temporal-graph convolutional network for p3-based video target detection. *Frontiers in Neuroscience*, *16*, 913027.

Lyu, X., Li, X., Dang, D., Dou, H., Wang, K., & Lou, A. (2022). Unmanned aerial vehicle (UAV) remote sensing in grassland ecosystem monitoring: A systematic review. *Remote Sensing*, *14*(5), 1096.

Mansfield, C. E. (2024). Decoding the recognition of occluded objects in the human brain. Ph.D. thesis. University of East Anglia.

Martinez-Trujillo, J. (2022). Visual attention in the prefrontal cortex. *Annual Review of Vision Science*, *8*(1), 407–425.

Mehler, B., Reimer, B., & Dusek, J. A. (2011). Mit agelab delayed digit recall task (n-back). *Cambridge, MA: Massachusetts Institute of Technology*, *17*, 33.

Min, B.-K., & Park, H.-J. (2010). Task-related modulation of anterior theta and posterior alpha EEG reflects top-down preparation. *BMC Neuroscience*, *11*, 1–8.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion* (pp. 69–84).

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Eeg signal classification for bci applications by wavelets and interval type-2 fuzzy logic systems. *Expert Systems with Applications*, *42*(9), 4370–4380.

Pascual-Marqui, R. D. et al. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): Technical details. *Methods and Findings in Experimental and Clinical Pharmacology*, *24*(Suppl D), 5–12.

Peyré, G., Cuturi, M. et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, *11*(5–6), 355–607.

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197.

Pogarell, O., Padberg, F., Karch, S., Segmiller, F., Juckel, G., Mulert, C., Hegerl, U., Tatsch, K., & Koch, W. (2011). Dopaminergic mechanisms of target detection-p300 event related potential and striatal dopamine. *Psychiatry Research: Neuroimaging*, *194*(3), 212–218.

Polich, J. (2007). Updating p300: An integrative theory of p3a and p3b. *Clinical neurophysiology*, *118*(10), 2128–2148.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z., & Wu, Q. M. J. (2022). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4051–4070.

Prabha, A. J., & Bhargavi, R. (2020). Predictive model for dyslexia from fixations and saccadic eye movement events. *Computer Methods and Programs in Biomedicine*, *195*, 105538.

Rahmani, M., Mohajelin, F., Khaleghi, N., Sheykhivand, S., & Danishvar, S. (2024). An automatic lie detection model using EEG signals based on the combination of type 2 fuzzy sets and deep graph convolutional networks. *Sensors*, *24*(11), 3598.

Rivet, B., Souloumiac, A., Attina, V., & Gibert, G. (2009). XDAWN algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Transactions on Biomedical Engineering*, *56*(8), 2035–2043.

Roy, A., Svensson, F. P., Mazeh, A., & Kocsis, B. (2017). Prefrontal-hippocampal coupling by theta rhythm and by 2–5 hz oscillation in the delta band: The role of the nucleus reuniens of the thalamus. *Brain Structure and Function*, *222*(6), 2819–2830.

Sajda, P., Gerson, A., & Parra, L. (2003). High-throughput image search via single-trial event detection in a rapid serial visual presentation task. In *First International IEEE EMBS conference on neural engineering, 2003. conference proceedings.* (pp. 7–10). IEEE.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Santamaria-Vazquez, E., Martinez-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.

Sauseng, P., Klimesch, W., Stadler, W., Schabus, M., Doppelmayr, M., Hanslmayr, S., Gruber, W. R., & Birbaumer, N. (2005). A shift of visual spatial attention is selectively associated with human EEG alpha activity. *European journal of neuroscience*, *22*(11), 2917–2926.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, *38*(11), 5391–5420.

Senoussi, M., Moreland, J. C., Busch, N. A., & Dugué, L. (2019). Attention explores space periodically at the theta frequency. *Journal of Vision*, *19*(5), 22–22.

Shi, J., Bi, L., Xu, X., Feleke, A. G., & Fei, W. (2024). Low-quality video target detection based on EEG signal using eye movement alignment. *Cyborg and Bionic Systems*, *5*, 0121.

Song, X., Yan, B., Tong, L., Shu, J., & Zeng, Y. (2020). Asynchronous video target detection based on single-trial EEG signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(9), 1931–1943.

Song, X., Zeng, Y., Tong, L., Shu, J., Li, H., & Yan, B. (2021). Neural mechanism for dynamic distractor processing during video target detection: Insights from time-varying networks in the cerebral cortex. *Brain Research*, *1765*, 147502.

Struye, J., & Latré, S. (2020). Hierarchical temporal memory and recurrent neural networks for time series prediction: An empirical validation and reduction to multilayer perceptrons. *Neurocomputing*, *396*, 291–301.

Uleru, G.-I., Hulea, M., & Manta, V.-I. (2022). Using hebbian learning for training spiking neural networks to control fingers of robotic hands. *International Journal of Humanoid Robotics*, *19*(06), 2250024.

Wang, J., Bi, L., Fei, W., Xu, X., Liu, A., Mo, L., & Feleke, A. G. (2024). Neural correlate and movement decoding of simultaneous-and-sequential bimanual movements using EEG signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Wang, P., Jiang, A., Liu, X., Shang, J., & Zhang, L. (2018). Lstm-based eeg classification in motor imagery tasks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(11), 2086–2095.

Wang, R., Liu, Y., Shi, J., Peng, B., Fei, W., & Bi, L. (2022). Sound target detection under noisy environment using brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 229–237.

Wang, X., Yao, F., Li, A., Xu, Z., Ding, L., Yang, X., Zhong, G., & Wang, S. (2023). Dronenet: Rescue drone-view object detection. *Drones*, *7*(7), 441.

Wendling, F., Koksal-Ersoz, E., Al-Harrach, M., Yochum, M., Merlet, I., Ruffini, G., Bartolomei, F., & Benquet, P. (2024). Multiscale neuro-inspired models for interpretation of EEG signals in patients with epilepsy. *Clinical Neurophysiology*, *161*, 198–210.

Williams, C. C. (2020). Looking for your keys: The interaction of attention, memory, and eye movements in visual search. In *Psychology of learning and motivation* (pp. 195–229). Elsevier (*vol. 73*).

Wu, Y., & He, K. (2018). Group normalization. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19).

Xia, X., Guo, Y., Wang, Y., Yang, Y., Shi, Y., & Men, H. (2024). Advancing cross-subject olfactory EEG recognition: a novel framework for collaborative multimodal learning between human-machine. *Expert Systems with Applications*, *250*, 123972.

Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875.

Yuan, Z., Zhou, Q., Wang, B., Zhang, Q., Yang, Y., Zhao, Y., Guo, Y., Zhou, J., & Wang, C. (2024). Psaeegnet: Pyramid squeeze attention mechanism-based cnn for single-trial eeg classification in rsvp task. *Frontiers in Human Neuroscience*, *18*, 1385360.

Yun, W. J., Park, S., Kim, J., Shin, M., Jung, S., Mohaisen, D. A., & Kim, J.-H. (2022). Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control. *IEEE Transactions on Industrial Informatics*, *18*(10), 7086–7096.

Zang, B., Lin, Y., Liu, Z., & Gao, X. (2021). A deep learning method for single-trial EEG classification in RSVP task based on spatiotemporal features of ERPs. *Journal of Neural Engineering*, *18*(4), 0460c8.

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