

# Beyond the Gaze: Peripheral Vision-Aware Visual Detection Failures Recognition through LLM-Based Fixation Coordinate-Sensitive Analysis

Zhimin Li, Fan Li, Gangyan Xu, *Member, IEEE*, Donglin Li

**Abstract**—Visual detection failures are a critical challenge in air traffic control (ATC), where undetected alerts can compromise operational safety and decision-making. Previous studies have primarily assessed detection failures through target fixation patterns, yet this method struggles to identify the more complex “look-but-fail-to-see” and “see-without-looking” scenarios. This underscores the necessity of exploring peripheral vision mechanisms, where dynamic tracking trajectories could better capture the scope of visual attention. Therefore, this study proposes a classification framework for visual detection by integrating peripheral vision tracking and human attentional states, including detection failures such as peripheral vision neglect and look-but-fail-to-see errors. A hierarchical detection failure recognition framework specific to the ATC settings is further developed and validated through an ATC simulation experiment. The framework first employs an Adaptive Symbolic Alert Detection method to identify and annotate ATC-specific alert regions with spatiotemporal uncertainty (achieving 95.24% precision), followed by LLM-based evaluation of operators’ visual attention to these regions to intelligently assign classification labels. Additionally, we introduce a fixation coordinate-sensitive multi-domain feature set that captures spatiotemporal and frequency-domain characteristics across detection types, achieving 93.13% four-class classification accuracy, outperforming traditional feature sets (83.69%) and both single- and dual-domain features (ranging from 76.82% to 90.11% accuracy). These findings demonstrate that our framework effectively captures a broader and structured range of visual detection failures, providing critical insights to improve the reliability of alert detection in ATC and the design of an intelligent human-centered ATC support system.

**Index Terms**—Visual detection, peripheral vision, look-but-fail-to-see error, eye movements, air traffic control.

## I. INTRODUCTION

**C**ONFLICT detection and resolution are vital tasks for air traffic controllers (ATCOs) [1]. With increasing air traffic density and the advancement of automation technology, controllers are required to monitor multiple digital displays

The study is partially supported by The Hong Kong Polytechnic University Research Centre Data Science AI (P0042711), National Natural Science Foundation of China (NSFC Project No. 52405295), and a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 25233824 for ECS project funded in 2024/25 Exercise). (Corresponding author: Fan Li.)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by PolyU Institutional Review Board (PROVIDED under Application No. HSEARS20240725004, and performed in line with the Declaration of Helsinki).

Zhimin Li, Fan Li, Gangyan Xu, and Donglin Li are with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: zhimin.li@connect.polyu.hk; fan-5.li@polyu.edu.hk; gangyan.xu@polyu.edu.hk; donglin.li@connect.polyu.hk).

[2]. This can lead to visual attention overload and detection failures of potential conflicts, with perceptual failures being the most common [3], [4]. Shorrock’s study, which included interviews with 28 UK area controllers and a review of 48 loss of separation incidents over three years, found that visual perceptual failures (e.g., no visual or late visual detection) accounted for nearly 70% of all perceptual errors [3]. In 2022, the U.S. Federal Aviation Administration reported 1732 runway incursions that were probably caused by detection failures to critical visual cues. It was reported that ATCOs fail to report around 40% of alarms, and 70% fail to detect critical stimuli on runways [5], [6]. Furthermore, detection failures account for 50% of failures in safety risk perception [7]. Since visual detection failures are an unavoidable byproduct of human behavior in dynamic contexts, it is critical to promptly and accurately identify these failures, as ATCOs are responsible for thousands of lives.

When evaluating visual detection failures, many studies rely on target fixations to measure whether an individual has detected the target [8]. For instance, when examining visual field defects, inattentive blindness (IB)—the failure to detect unexpected but salient stimuli in the visual field—is the most frequently observed visual detection failure [8], [9]. The conventional criterion for IB is the absence of gaze on the stimulus [10], [11]. However, not fixating on a target does not necessarily mean it was not noticed [12]. This approach overlooks more complex scenarios, such as “look-but-fail-to-see” errors and “see-without-looking” phenomena—cases that reflect distinct attentional mechanisms. “Look-but-fail-to-see” errors occur when an individual fixates on a stimulus but fails to cognitively process its existence, resulting in no response [13]. This can be attributed to over-attention to a specific cognitive task, leading to a lack of situational awareness of unexpected stimuli despite direct visual engagement [14], [15]. On the other hand, “see-without-looking” describes a situation where a stimulus is detected despite the absence of direct fixation, primarily driven by covert attention and peripheral vision processing [16], [17]. This phenomenon is known as pre-attentive or automatic detection, where peripheral vision facilitates the unconscious processing of unexpected stimuli and triggers reflexive responses [18], [19]. Neurophysiological evidence further supports this mechanism, revealing distinct brain responses when peripheral stimuli are detected without direct foveal fixation [12]. Considering the above two complex scenarios would provide a deeper understanding of detection failures in visual tasks, particularly in high-stakes operational

environments such as ATC.

Therefore, this study aims to propose a novel peripheral vision-aware visual detection classification to recognize involved detection failures. This classification encompasses two aspects: the gaze status on the stimulus and the individual's corresponding behavioral response. Specifically, when the gaze is directed at a target, a lack of response in this context suggests a “look but fail to see” error [2], while a corresponding response denotes “effective detection” [20]. Conversely, if the gaze is not directed at the target but a correct response occurs, it implies “pre-attentive detection” via peripheral vision [12]. Failure to respond without gaze indicates “peripheral vision neglect.” Despite this categorization, the understanding of the visual properties underlying these four detection types is limited, and no empirical data are available to support these distinctions and accurately recognize the involved detection failures.

Accurately recognizing the four different visual detection types via peripheral vision-aware visual detection classification in an ATC environment presents significant challenges. On the one hand, a major challenge is the inherent spatiotemporal uncertainty of alerts in dynamic natural scenes. This uncertainty complicates the detection of alerts and the verification of whether a controller's gaze has landed on them. For instance, controllers must monitor various dynamic displays, such as radar and auxiliary displays, for weather and airport information [2]. On the primary radar display, each aircraft is represented by a blip and a label containing information like callsign, flight level, and destination code, with short-term conflict alerts causing the label to turn red [3]. The spatiotemporal uncertainty of alert positions and timings makes it impractical to use traditional pre-defined Areas of Interest (AOIs) to determine if a controller's gaze has landed on an alert [21]. Thus, eye-tracking object detection approaches using YOLO (You Only Look Once) were recently proposed in [22] and [23], but they have limitations. While the YOLO system is pre-trained to recognize common objects such as vehicles and animals, the ATC radar interface is unique and not composed of typical physical entities, requiring a more tailored alert recognition method [3]. Additionally, verifying whether a controller's gaze has landed on the alert after alert detection remains time-consuming and prone to human error. In recent years, the rise of generative AI and vision-language models like CLIP offers promising solutions for overcoming challenges in complex scenarios, such as task-based zero-shot transfer for distracted driver activity recognition, highlighting their potential in intelligent behavior evaluation [24], [25]. Large Language Model (LLM) has been applied to automate annotation tasks, particularly through natural language understanding and semantic labeling, where human verification is combined with LLM efficiency to improve accuracy in complex datasets [26], [27]. Inspired by this, applying the image processing capabilities of LLMs to assess human visual attention on alerts offers a novel approach to enhancing alert detection processes [23], [28].

On the other hand, whether gaze trajectory can adequately reflect peripheral vision processing has not been explored [29]. Traditional eye-tracking features like blink rate, pupil size,

and fixation duration, while extensively used to assess visual attention in the central visual field [30]–[32], tend to oversimplify the complexities of human visual processing, especially regarding peripheral awareness [33]. Peripheral vision plays a critical role in detecting stimuli outside the direct line of sight, relying on broader visual tracking mechanisms that require different visual strategies, such as scanning movements and subconscious tracking, which are not reflected in traditional eye-tracking features [16], [18], [34]. These broader patterns may involve rhythmic scanning and temporal shifts critical for detecting peripheral stimuli. As indicated in [35], features integrating graph theory provide a more precise interpretation of the visual tracking dynamics. Although meeting the prerequisite of a graph neural network is challenging due to non-coinciding fixations in visual trajectories, its concept inspires us to fully explore the multi-layer information of visual trajectories. For instance, by employing spatial analysis, Fourier transform, and time-variant autocorrelation—methods that span multiple levels of signal processing, from spatial distribution to frequency analysis and temporal dynamics—we can uncover novel multidimensional embedding features [36], [37].

To address the aforementioned challenges, this paper proposes a hierarchical peripheral vision-aware detection failure recognition framework to assess human attentional states toward ATC-specific alerts. By automating the recognition of detection failures, the proposed framework can be integrated into air traffic management systems to provide effective feedback on controller attention states. This capability not only improves the reliability of alert detection in ATC, but also establishes a foundation for the development of an intelligent, human-centered ATC support system. The contributions of this study are summarized as follows:

1. This paper proposes a visual detection classification that accounts for peripheral vision tracking and human cognitive states and further develops a hierarchical framework specific to the ATC context to recognize the proposed detection failures. Validated through an ATC simulation experiment, the framework illustrates distinct visual tracking patterns across four detection categories, highlighting the diversity and systematic nature of detection failures and providing critical insights for improving alert detection performance.

2. This paper introduces an Adaptive Symbolic Alert Detection method to assess location-specific alerts on ATC radar interfaces. This innovative method utilizes color segmentation, morphological transformations, and text interpretation to accurately identify and annotate alerts, adapting to the dynamic and time-sensitive nature of alerts in the air traffic monitoring system.

3. Addressing the challenge of time-consuming and error-prone manual labeling of visual attention to alerts, this study demonstrates the potential of GPT-4V with LLM to ascertain controllers' visual attention engagement with alerts, offering a more efficient and intelligent approach to alert detection assessment, enhancing convenience and reliability in air traffic monitoring and accident prevention.

4. A novel fixation coordinate-sensitive multi-domain feature set is developed, analyzing gaze trajectories across spatial,

frequency, and time domains. These features outperform traditional ones in recognizing peripheral vision-aware detection failures and adapting to the ATC system through diverse parameter configurations, improving safety and efficiency in transportation management.

This paper is organized as follows. In Section II, the experimental setup is outlined. The proposed methodology is explained in Section III. The results are detailed in Section IV, and the discussions are presented in Section V. Finally, the conclusions are provided in Section VI.

## II. EXPERIMENTAL SETUP

### A. Participants

This study evaluated the hierarchical framework via an ATC simulation at the Hong Kong Polytechnic University with 38 participants (age  $M = 23$  yr,  $SD = 13.60$ ,  $Max = 38$  yr,  $Min = 19$  yr), including 27 pre-professionals with ATC qualification certificates and 11 members from VATSIM (Virtual Air Traffic Simulation Network) Hong Kong, a global online community for realistic ATC simulation, all with over 1 year of simulated ATC training and simulated ATC experience, ensuring ATC proficiency and consistency. All participants had normal or corrected-to-normal vision. To ensure data security, all visual data in ATC simulation experiment were anonymized, with only relevant segments (e.g., eye movement trajectories and specific alert regions) processed by GPT-4V. Participants provided informed consent, and the study was approved by the IRB (HSEARS20211117002), adhering to ethical standards and guidelines.

### B. Apparatus

In this study's methodological framework, a desktop computer with a 27-inch monitor (1920\*1080 pixels) was paired with a Gazepoint 3 (GP3) eye-tracker to collect detailed eye movement data, as shown in Fig. 1. The eye-tracker software operates on a separate laptop connected to the monitor via HDMI. The GP3, an integrated desktop eye-tracking system, records various eye movement metrics at a sampling rate of 60 Hz, using binocular tracking within an operating distance of 50 cm to 80 cm from the monitor. The system's spatial accuracy ranges from  $0.5^\circ$  to  $1^\circ$  of visual angle, with a spatial resolution of  $0.1^\circ$  RMS. Calibration was performed using a 9-point method, with successful calibration requiring accurate fixation on each point. Calibration was checked regularly, and recalibration was performed if discrepancies were detected. Additionally, EEG data was collected using a synchronized headset to support future gaze-neural correlation analyses in ATC detection failures, with pre-experiment comfort assessments confirming no significant discomfort and ensuring data quality [38], [39].

### C. Experiment procedures

The experimental procedure was designed to simulate the complex environment characteristic of ATC operations. The experimental interface, depicted in Fig. 2, comprises a radar-map interface from the Euroscope platform and a simplified

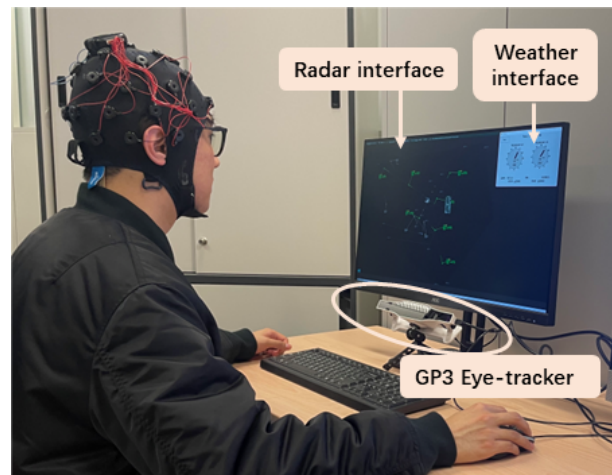


Fig. 1: Experimental setup showing a 27-inch monitor (1920×1080 pixels) paired with a Gazepoint 3 eye-tracker, recording eye movements data.

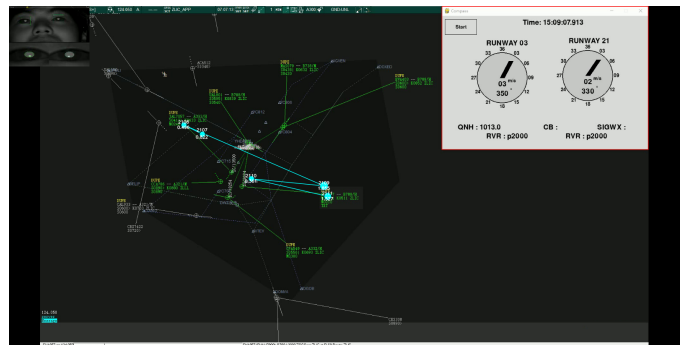


Fig. 2: Experimental interface for the ATC multi-tasking scenario, featuring a radar-map display from the Euroscope platform (center and left) and a simplified weather interface (upper right), with blue gaze trajectories recorded by an eye tracker overlaid on the user interface.

weather interface, as informed by prior work on ATC human factors [3]. Participants engaged in a 40-minute multi-tasking scenario, managing the safe arrival of 17 aircraft, each requiring specific descent and approach instructions. This scenario encompassed radar surveillance, issuing clearances, weather monitoring, and radiotelephony communication with a pre-arranged pseudo-pilot, an experienced ATC professional who simulated realistic flight scenarios to create an immersive operational environment [3]. Additionally, the scenario incorporated randomly appearing aircraft, termed fly-over flights, crossing the radar airspace—a common feature in real ATC operations. These flights, while not requiring direct participant intervention, increased cognitive load by necessitating sustained situational awareness in a busy airspace. Overall, the experimental duration was designed to simulate the sustained cognitive load experienced in practical ATC operations while ensuring sufficient data collection for robust analysis. It aligns with similar studies, such as [40], which employed around 40-minute simulated driving tasks to simulate high-stress environments and collect robust data.

During each 40-minute ATC simulation session, a minimum of 13 alerts per participant ( $M=16$ ,  $SD=2$ ) were randomly triggered, with additional alerts varying based on the simulation's evolving traffic conditions, to simulate comprehensive unpredictability of alert presentation and operational demands in multitasking conditions, aligning with previous studies (e.g., [41] used probabilistic alerts; [42] employed variable alert scenarios). The distribution of alerts across participants is detailed in Appendix A. Post-experiment feedback validated the design's workload appropriateness, with participants reporting a workload of 5–7 on a 1–9 scale, indicating a challenging yet manageable cognitive demand. Throughout each 40-minute session, a total of 13 pre-programmed alerts were randomly triggered to evaluate participants' ability to detect and respond to sudden visual stimuli amidst concurrent tasks. Each alert, defined as a single data point, lasted 5 seconds and appeared individually at random intervals and locations. These alerts were initiated by the pseudo-pilot, who randomly selected one fly-over flight to simulate an emergency by activating the 7700-emergency squawk code—an ATC standard for distress signals [43]. These brief, attention-grabbing signals, illustrated in Fig. 3, caused the label of an aircraft to turn red, indicating potential conflicts. This approach ensured that the timing and location of each alert were unpredictable, reflecting the dynamic and complex nature of real-world ATC operations. Participants were instructed to monitor these alerts peripherally while focusing on their primary tasks. Upon noticing an alert, they were required to press the “A” key, providing a measurable indicator of their attentional engagement. This setup ensured a dynamic assessment of attention allocation under realistic, unpredictable conditions. Additionally, based on the current control situation, they could engage with the pseudo-pilot of the alerted aircraft, inquiring about flight intentions to assess priority levels and ensure air traffic safety. The alerts automatically resolved after 5 seconds, preventing additional cognitive load from interfering with their ongoing multi-tasking responsibilities.

In our experiment, we focused on the radar display as the primary information source due to its critical role in ATCOs' workflow. Radar alerts, such as those triggered by aircraft proximity violations, are the most immediate and high-priority tasks for controllers [44], [45]. We examined responses to single radar alerts, a rare but critical aspect of ATC operations [46]. This aligns with prior research on isolating critical events to identify distinct visual and behavioral patterns [43], [44], aiming to establish a foundational understanding of detection failures and their eye-tracking characteristics.

#### D. Gaze data collection

In this study, gaze data were collected within a 7-second window following the onset of each pre-programmed alert, comprising a 5-second alert presentation duration and a 2-second post-alert response buffer. This 5-second alert duration ensures consistent reaction times [47], while also optimizing response effectiveness in emergencies as noted by [48]. However, in multitasking scenarios, the process from perceiving an alert to responding often involves attentional allocation or

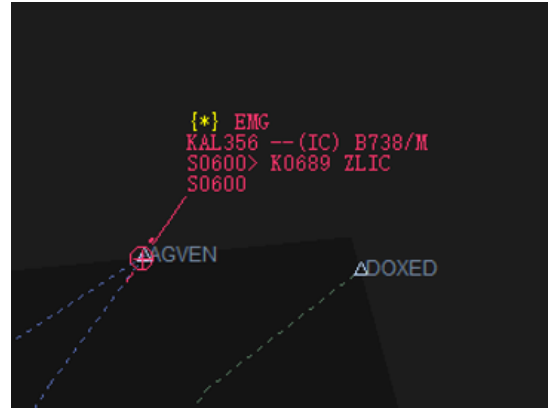


Fig. 3: Example of a pre-programmed alert on the radar interface, showing a red aircraft label indicating a potential conflict during a 40-minute ATC simulation.

task-switching demands, resulting in a temporal lag, which necessitates the 2-second response buffer to capture delayed responses. This buffer, further supported by [49], accounts for residual attention processing, as controllers may require additional time to process and respond. As a result, the 7-second window comprehensively captures the complete gaze-behavior sequence, enabling accurate detection and response assessment in alignment with our framework for classifying detection failures. The timeline of this process is illustrated in Fig. 4, which depicts the 7-second window, including the alert presentation duration, response buffer, and gaze-behavior sequence, while highlighting potential delayed responses. Consequently, each 7-second interval is treated as an individual sample for detailed behavioral analysis.

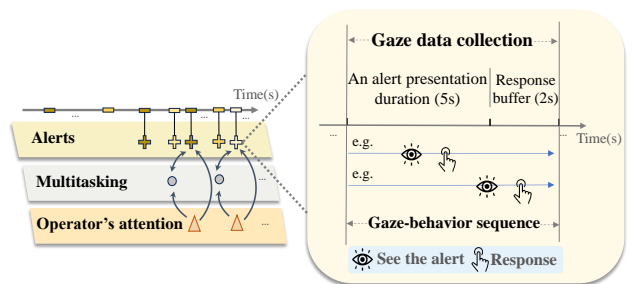


Fig. 4: Timeline of the 7-second gaze data collection window, comprising a 5-second alert presentation duration and a 2-second response buffer. The gaze-behavior sequence below illustrates alert detection and response, highlighting potential delayed reactions in a multitasking context.

### III. METHODOLOGY

This study first classifies visual detection types by integrating peripheral vision tracking and human attentional states derived from human fixations and behavior. A **Hierarchical Peripheral Vision-aware Detection Failures Recognition Framework** specific to ATC alerts is further developed to recognize the proposed detection failures, as depicted in Fig. 5, and validated through an ATC simulation experiment yielding

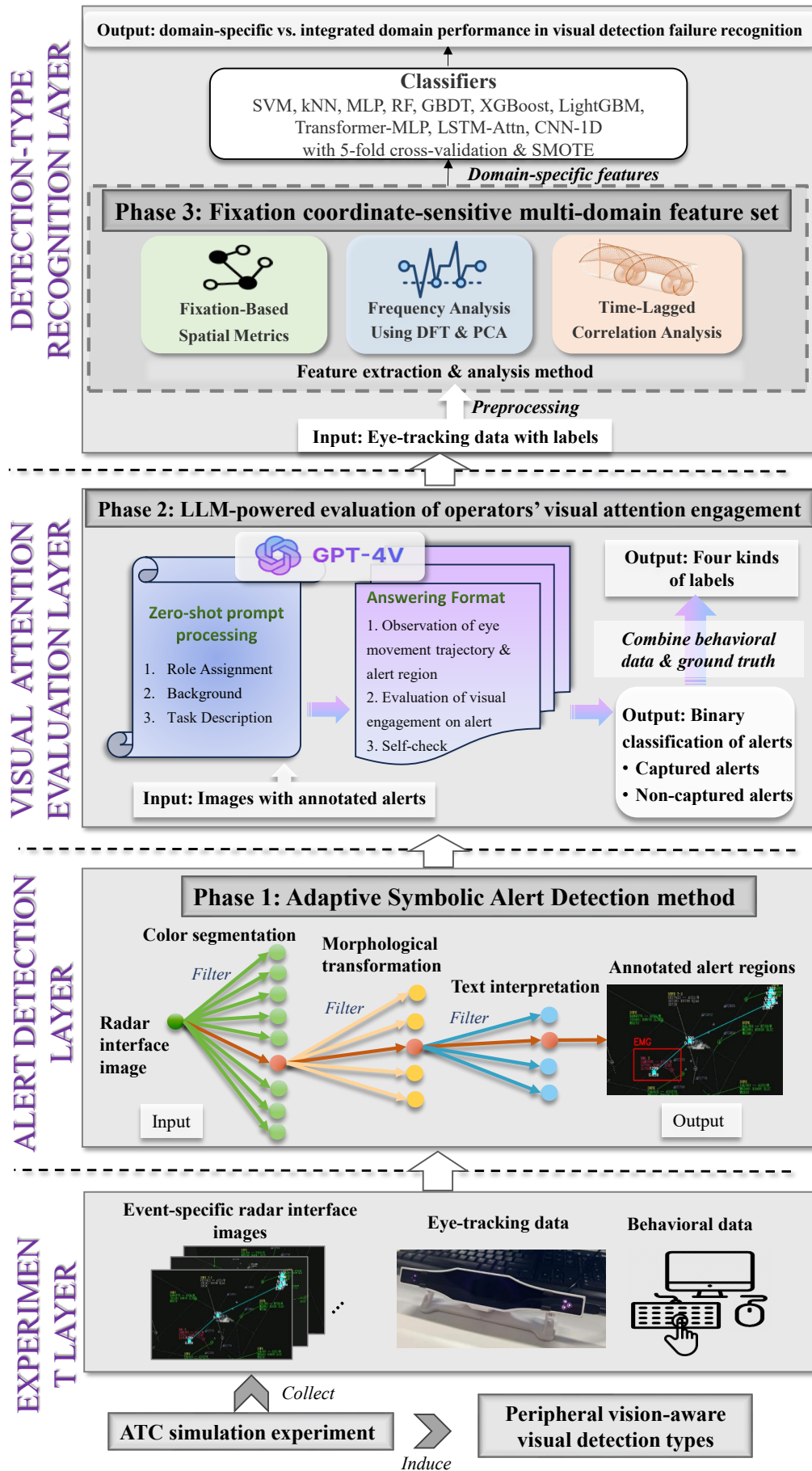


Fig. 5: Hierarchical peripheral vision-aware detection failure recognition framework for ATC-specific system, integrating ATC simulation, alert detection and annotation, LLM-based visual attention evaluation, and multi-domain feature analysis.

empirical data. The framework comprises three phases: (1) an Adaptive Symbolic Alert Detection method is introduced to identify alert regions with spatiotemporal uncertainty on the ATC radar interface; (2) a GPT-4V-powered evaluation of visual attention engagement to alerts is employed to facilitate labeling samples into four detection types; and (3) a Fixation Coordinate-Sensitive Multi-Domain Feature Set is proposed to analyze fixation dynamics to classify the four detection types, comparing domain-specific and integrated feature performance.

#### A. Peripheral vision-aware visual detection classification

The detection types in this study are categorized based on a combination of gaze behavior and participant responses, as illustrated in Fig. 6. The figure outlines a decision-making process starting with an alert's appearance and branching into different outcomes based on whether the participant's gaze was directed at the alert and their timely response. These detection outcomes serve as the ground truth for evaluating detection methods, ensuring consistency and a well-defined standard for classification tasks.

Firstly, peripheral vision processing is characterized by the appearance of an alert within the participant's peripheral vision, a critical phase that determines whether the alert will be subsequently attended to. If the participant responds to the alert without directly gazing at it, the incident is classified as **Pre-Attentive Detection** [18], [19]. This indicates that peripheral vision was effective in detecting the alert, even though the operator's focus remained on other tasks, possibly delaying their attention to the alert itself. This classification will be essential in evaluating detection methods that leverage peripheral awareness, as it establishes the baseline for identifying alerts that were recognized without direct gaze. In contrast, **Effective Detection** occurs when the participant not only shifts their gaze toward the alert but also responds within the allotted time [20]. This suggests that the participant visually processed and cognitively acknowledged the alert, considering its implications for the broader radar environment and aircraft management. The Effective Detection category provides a critical benchmark for assessing methods designed to ensure timely and appropriate responses to alerts, serving as a reference point for successful alert recognition and reaction. Another critical classification is the **Look-but-Fail-to-See (LBFTS) Error** [13], [50]. This occurs when the participant's gaze does land on the alert, yet they fail to respond within the designated time frame. Such an outcome indicates that while the alert was visually perceived, it was not cognitively processed in a manner that prompted an appropriate reaction [17]. This error highlights a breakdown in the link between visual attention and subsequent cognitive action. By including LBFTS Errors in our classification, we provide crucial ground truth for identifying situations where visual perception does not translate into an effective response. This allows for the analysis of cognitive processing failures in detection tasks. Lastly, **Peripheral Vision Neglect** is observed when the participant neither gazes at the alert nor responds within the time limit, indicating a failure to detect alerts through peripheral

vision. This type is essential to assess, as peripheral vision is crucial for detecting unexpected events and maintaining situational awareness without direct fixation, as highlighted in [17]. This classification will be used as ground truth to evaluate the worst-case scenarios in detection methods, where both visual attention and peripheral awareness fail, offering insights into the limitations and challenges in alert detection.

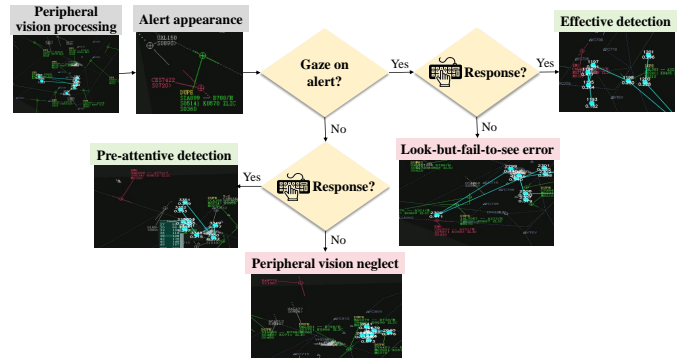


Fig. 6: Decision tree classifying peripheral vision-aware visual detection types (Effective Detection, Peripheral Vision Neglect, LBFTS Error, Pre-attentive Detection) based on gaze and response behavior in ATC alert scenarios.

#### B. Phase 1: Adaptive Symbolic Alert Detection Method

This study implements a robust image processing framework designed to enhance the detection and annotation of critical alerts on ATC radar interfaces. Each step in our methodology is precisely calibrated to optimize accuracy and efficiency, drawing from advanced computational techniques to process images at a resolution of 512x288 pixels. The image processing was implemented using Python 3.11, with the following libraries employed: OpenCV (version 4.9.0.80) for image processing tasks such as color segmentation and morphological operations, and NumPy (version 1.23.5) for handling array operations and matrix manipulations. Phase 1, shown in Fig. 5, outlines the steps of the adaptive symbolic alert detection method. The term "Adaptive Symbolic" reflects the method's ability to dynamically process and integrate various types of symbolic information—such as colors, contours, and text—each representing different aspects of an alert. By adapting to the specific characteristics of these symbolic elements, the approach can effectively detect and annotate alerts in a manner that is both contextually aware and highly precise, ensuring that the system responds appropriately to the complexity of the information it encounters.

The first stage involves effective color segmentation techniques to isolate alerts based on color cues, specifically red, commonly used in ATC systems to denote urgency. The process begins with converting the input images to the Hue, Saturation, Value (HSV) color space to better distinguish color variations under different lighting conditions [51]. Specific HSV ranges are targeted:

- Lower red range: HSV values of (0, 120, 70) to (10, 255, 255)

- Upper red range: HSV values of (170, 120, 70) to (180, 255, 255)

These thresholds are chosen based on empirical data indicating their effectiveness in capturing a wide spectrum of red tones displayed on diverse ATC interface technologies. Binary masks are then generated to identify areas of the image falling within these red ranges, using OpenCV's `cv2.inRange()` function.

The second stage, morphological transformations, refines the segmented regions to ensure accurate detection. This involves applying a series of filters:

- A closing operation with a 7x7 pixel kernel to close small holes and connect broken regions within the alert areas, ensuring continuity.
- An opening operation to remove minor noise points, followed by dilation to enhance the visibility and distinctness of the contours.

These transformations are crucial in preparing the segmented areas for precise contour detection, enabling the identification of critical elements such as circular symbols (representing aircraft icons), associated text labels (displaying flight information and status), and connecting lines. The processed image undergoes contour analysis to isolate these components, with the Hough Line Transformation employed to detect any linear elements connecting the symbols and labels.

The third stage, text interpretation, integrates all identified elements into a cohesive alert region. In this step, the detected contours and lines are merged into a single bounding rectangle that encompasses the entire alert, including both the circular symbols and the associated text labels. This unified region is then annotated with "EMG" above it. The annotation serves as a clear and precise interpretation of the alert's nature, ensuring that it is properly highlighted and ready for further analysis by LLMs. This comprehensive process effectively prepares the images to assess alert visibility and attention tracking in ATC settings.

This method effectively segments, refines, and integrates critical alert elements—such as aircraft icons and flight information—into a unified region. By accurately detecting and annotating these components, the system ensures clear and reliable identification of critical alerts, optimizing the images for further analysis in ATC settings.

### C. Phase 2: GPT-4V-powered evaluation of visual attention engagement

This study tackles the challenge of manually verifying a controller's attention to alerts by utilizing GPT-4V, a visual extension of the GPT-4 model, in conjunction with LLMs. Phase 2, depicted in Fig. 5, outlines the refining summary of LLM-powered evaluation of operators' visual attention engagement on alerts. Specifically, the GPT-4o model was selected for its advanced capabilities in visual and language processing, deemed suitable for handling the complex visual inputs typical in ATC scenarios. The model provided a nuanced and context-aware analysis of controllers' visual attentional engagement with alerts, as detailed in Algorithm 1.

### Algorithm 1 Evaluating visual attention engagement using GPT-4V

**Require:** Image dataset directory, OpenAI API key

**Ensure:** Responses are saved to an Excel file

- 1: Initialize variables:
- 2: `openai.api_key` ← Retrieve from environment variable or set manually.
- 3: `folder_path` ← Path to the directory containing images.
- 4: `image_paths` ← List all image files in the specified directory.
- 5: `models` ← ['gpt-4o'].
- 6: Define prompt for analyzing ATC images:
  - Role Assignment:

```
Assume the role of a researcher analyzing eye
movement trajectories in ATC control tasks.
```

- Background:

```
Analyze images from a simulated ATC interface
showing various detailed information
with ['status indicators'], ['flight
information'], ['emergency alerts'],
['emergency highlighting'], ['eye movement
visualization'].
```

- Task Description:

```
Determine if eye movement trajectories focus
on areas highlighted for emergencies. Verify
the presence of blue trajectory markers within
red emergency zones, classifying observations
as 'Captured alert' or 'Non-captured alert'
based on their alignment.
```

- 7: Process each image using the defined models:
- 8: **for** each `image_path` in `image_paths` **do**
- 9:     Convert image to base64 encoding
- 10:    **for** each model in `models` **do**
- 11:       Construct payload for API request
- 12:       Send POST request to OpenAI's API
- 13:       **if** response is successful **then**
- 14:          Parse and save response
- 15:       **else**
- 16:          Log error details
- 17:       **end if**
- 18:     **end for**
- 19: **end for**
- 20: Compile responses into a DataFrame
- 21: Save the DataFrame to an Excel file at the specified path

To generate initial binary labels, we employed zero-shot prompts tailored to the specific context of ATC interfaces [52]. Each event-specific image, annotated with alerts, was encoded into a suitable format and input into the GPT-4V model. The prompt structure defined the model's role, background context, and task objectives. It was instructed to act as a researcher analyzing eye movement trajectories in ATC tasks, specifically assessing whether blue trajectory markers (fixation trajectory) fell within red emergency zones (alert region). The task was to classify eye movement trajectories into Captured alert or Non-captured alert based on gaze engagement with the alert

region. Non-captured alerts were assigned when no fixations were detected within the designated alert regions, while their presence indicated a Captured alert.

This classification approach considers the distinction between central and peripheral vision, a key factor in assessing gaze engagement. Central vision, defined by the fovea, spans approximately  $2^\circ$  of visual angle for high-acuity perception, while peripheral vision extends beyond this range to  $90^\circ$ – $100^\circ$  eccentricity, with decreasing detail sensitivity [53], [54]. Rather than applying a fixed  $2^\circ$  range, we assess central vision engagement by determining if the fixation trajectory overlaps the alert region. This approach is grounded in three primary considerations: its theoretical basis in foveal dynamics, its alignment with ATC task demands, and its methodological efficacy. First, fixations, recorded as stabilized gaze positions (200–300 ms), dynamically reflect foveal focus [54]. The fixation trajectory embeds the  $2^\circ$  foveal span within real-time gaze behavior, aligning with the fovea’s natural targeting process without imposing a fixed angular limit. Second, since ATC alerts vary in size and position due to factors like aircraft labels and emergency trigger, the annotated alert regions are directly aligned with task-relevant targets, avoiding mismatches a fixed range might introduce [3]. Third, a  $2^\circ$  range demands pixel conversion and overlap assessment with a hypothetical circular zone per fixation point, increasing computational complexity and potential mismatch with variably sized alerts. Our approach relies on the direct spatial relationship between fixation trajectory and the alert region, providing a straightforward and efficient criterion.

To ensure the reliability of binary labels, the model sent payloads to the OpenAI API for parsing and analysis, with each image requiring approximately 750 tokens for processing and generating categorical labels based on prompt-based reasoning. This processing, involving batch analysis of eye movement trajectories collected post-data acquisition, results in delays typically ranging from 320 milliseconds to 1 second per request, ensuring data integrity. Moreover, to validate these labels, we aligned the LLM outputs with ground truth labels established through expert evaluations. Therefore, the final labels used for model classification were 100% consistent with the ground truth. This method is similar to that described in [55], where a natural prompt was used to generate initial LLM labels, which were then refined to align the outputs with ground truth labels. Alternative methods, such as traditional AOI settings and computer vision techniques, offer insights in some cases but have limitations in intelligently assessing visual attention dynamics in relation to alert regions in ATC scenes [56]. Thus, expert verification remains the most suitable comparison for this task. Finally, the binary labels (Captured/Non-captured alerts) were further refined by integrating behavioral response data to derive four visual detection types: Pre-attentive detection, Peripheral vision neglect, LBFTS error, and Effective detection. By integrating LLM outputs with ground truth verification, we minimized the need for extensive manual labeling while maintaining methodological rigor.

#### D. Phase 3: Fixation coordinate-sensitive multi-domain feature set

To comprehensively capture the nuanced patterns of visual tracking dynamics, we propose a novel fixation coordinate-sensitive multi-domain feature set. This approach integrates spatial, frequency, and time-based features derived from gaze and fixation coordinate data, leveraging the synergistic interplay among these domains to provide a robust representation of gaze behavior. As illustrated in Fig. 7, this section will detail how these dimensions are combined to form a comprehensive feature set for peripheral awareness detection analysis.

1) **Fixation-Based Spatial Metrics:** The analysis of gaze behavior in this study incorporates fixation-based spatial metrics, which are essential for quantifying how visual attention is distributed across the visual field. These metrics include gaze stability, fixation dispersion, the number of fixation clusters, and total scanpath length, each contributing to a comprehensive understanding of visual attention dynamics.

**Gaze stability** is assessed by calculating the standard deviation of positions in pixels of gaze points that belong to each fixation within a specified time window [34]. Mathematically, if  $(x_i, y_i)$  represents the coordinates of gaze points within a fixation, the gaze stability  $\sigma_{\text{gaze}}$  is given by:

$$\sigma_{\text{gaze}} = \sqrt{\frac{1}{N} \sum_{i=1}^N [(x_i - \bar{x})^2 + (y_i - \bar{y})^2]} \quad (1)$$

where  $\bar{x}$  and  $\bar{y}$  are the mean coordinates of the gaze points in the fixation. This metric reflects the consistency of gaze within a fixation, with lower values indicating more stable gaze and higher values suggesting greater variability.

**Fixation dispersion** is evaluated by calculating the root mean square of the distances from each fixation point to the average fixation position within a time window [57]. This metric provides insight into how spread out the fixation points are relative to their mean position. The fixation dispersion  $D_{\text{dispersion}}$  is calculated as:

$$D_{\text{dispersion}} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left[ \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2} \right]^2} \quad (2)$$

This measure helps to quantify how widely the participant’s attention is distributed during the task, with larger values indicating greater dispersion of fixation points.

The **number of fixation clusters** is determined using clustering algorithms such as k-means, which identify groups of fixation points that are spatially close. Each cluster represents a distinct area of interest that repeatedly attracts the participant’s gaze. The number of clusters  $C$  indicates how many different regions of interest were focused on during the task. A higher number of clusters may reflect more distributed attention across multiple areas, whereas fewer clusters suggest more focused attention.

Finally, **total scanpath length**  $L$  is calculated to evaluate the overall distance traveled by the gaze across all fixation

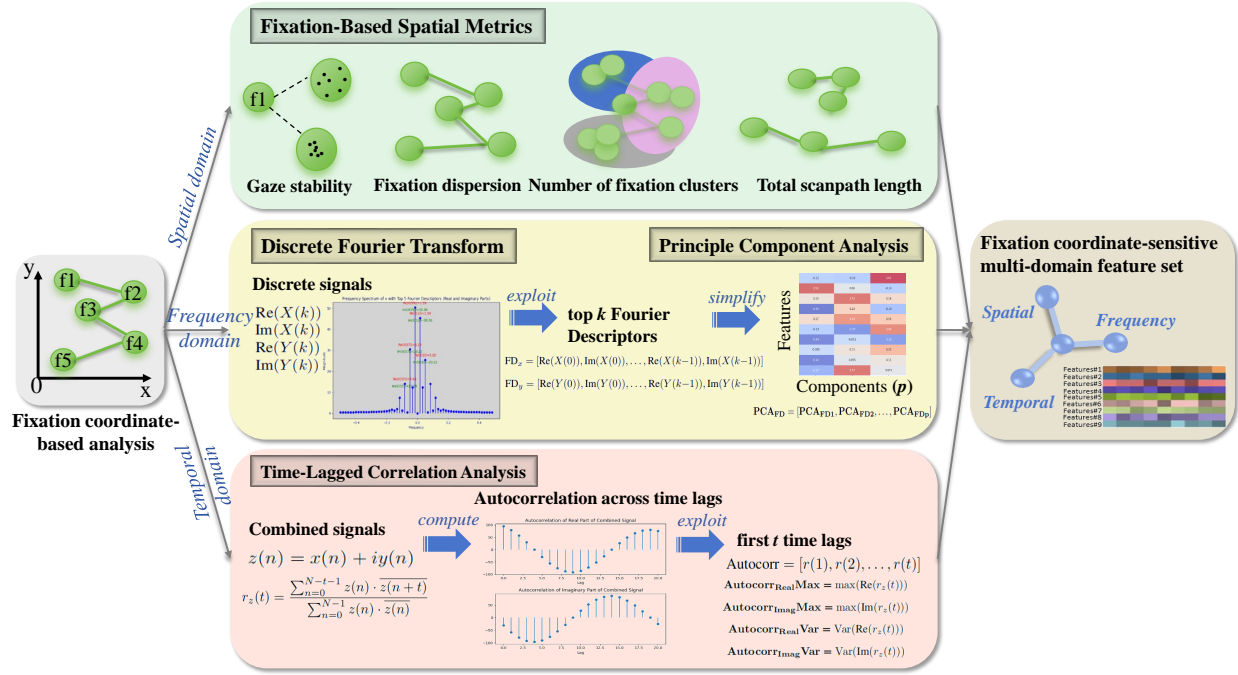


Fig. 7: Overview of the proposed fixation coordinate-sensitive multi-domain feature set, combining spatial metrics, frequency descriptors, and time-lagged autocorrelation for analyzing fixation trajectories in ATC alert detection tasks.

points [35]. It is a measure of the visual exploration strategy employed by the participant and is defined as:

$$L = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \quad (3)$$

A longer scanpath length suggests a more exhaustive search behavior, potentially indicating the need for more visual information gathering, while a shorter scanpath indicates a more efficient and focused scanning pattern.

Together, these fixation-based spatial metrics provide a detailed characterization of how participants allocate their visual attention across space, offering critical insights into the mechanisms that drive effective visual processing and peripheral awareness during task performance.

2) **Frequency Analysis Using Discrete Fourier Transform:** In the frequency domain, we analyze the horizontal ( $x$ ) and vertical ( $y$ ) coordinates of fixations separately, as each direction may exhibit distinct frequency characteristics. Separating them ensures that we capture their individual oscillatory patterns and periodic behaviors without interference from cross-dimensional interactions. The frequency domain features derived from Fourier Descriptors (FDs) and PCA. These methods allow us to extract meaningful characteristics from the gaze data and distinguish between different detection types by focusing on the spatial distribution and movement patterns of fixation points.

FDs provide a way to represent the shape of a signal in the frequency domain [58]. Given a sequence of fixation points represented as coordinates over time, the Fourier Transform is applied to convert these coordinates from the time domain to the frequency domain. This transformation is beneficial be-

cause it allows us to capture periodic patterns and oscillations in the gaze data.

Given a sequence of fixation points  $(x_i, y_i)$  where  $i = 0, 1, \dots, N - 1$ , we treat  $x$  and  $y$  as discrete signals. The Discrete Fourier Transform (DFT) is applied to both  $x$  and  $y$  to obtain their frequency domain representations. For a discrete signal  $x(n)$  with  $N$  samples, the DFT is defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i \frac{2\pi kn}{N}} \quad (4)$$

where  $X(k)$  is a complex number comprising a real part  $\text{Re}(X(k))$  and an imaginary part  $\text{Im}(X(k))$ :

$$\text{Re}(X(k)) = \sum_{n=0}^{N-1} x(n) \cdot \cos\left(\frac{2\pi kn}{N}\right) \quad (5)$$

$$\text{Im}(X(k)) = - \sum_{n=0}^{N-1} x(n) \cdot \sin\left(\frac{2\pi kn}{N}\right) \quad (6)$$

These components represent the cosine and sine components of the signal at each frequency  $k$ . The same approach applies to  $Y(k)$ .

Given the sampling frequency of 60 Hz and a time window of 5 seconds, the frequency range spans from -30 Hz to 30 Hz, with a frequency resolution of 0.2 Hz. We systematically vary the number of Fourier descriptors ( $n_{\text{descriptors}}$ ) to identify the optimal balance between capturing key oscillatory patterns and maintaining model generalization. Lower values may retain only coarse movement trends, while higher values introduce finer details that may not improve classification. This exploration ensures a compact yet informative representation of fixation dynamics for robust recognition. Thus, the feature

vector for each configuration is formed by concatenating the real and imaginary parts of the selected frequency components ( $n_{\text{descriptors}} = k$ ) for both  $x$  and  $y$  coordinates:

$$\text{FD}_x = [\text{Re}(X(0)), \text{Im}(X(0)), \dots, \text{Re}(X(k-1)), \text{Im}(X(k-1))] \quad (7)$$

$$\text{FD}_y = [\text{Re}(Y(0)), \text{Im}(Y(0)), \dots, \text{Re}(Y(k-1)), \text{Im}(Y(k-1))] \quad (8)$$

The combined feature vector for both coordinates is:

$$\text{FD} = [\text{FD}_x, \text{FD}_y] \quad (9)$$

This feature vector provides a detailed frequency-based representation of the fixation points, highlighting essential periodicities and shape characteristics in the eye movement data.

Given the dataset  $\text{FD}$  of shape  $(m, n)$ , where  $m$  is the number of samples and  $n$  is the number of features, PCA is applied to reduce dimensionality while preserving key variations [59]. It transforms the original descriptors into uncorrelated principal components—linear combinations of the original features—that capture the most significant data variations. The process involves:

1. **Standardization:** Subtract the mean of each feature and divide by its standard deviation to ensure all features contribute equally to the analysis.

2. **Covariance Matrix Computation:** Calculate the covariance matrix  $\Sigma$  to understand how the features vary together:

$$\Sigma = \frac{1}{m-1} \text{FD}^T \text{FD} \quad (10)$$

3. **Eigenvalue Decomposition:** Decompose the covariance matrix into eigenvalues and eigenvectors. The eigenvectors represent the directions of maximum variance, while the eigenvalues represent the magnitude of this variance.

4. **Principal Components Selection:** Initially, the top 3 eigenvectors corresponding to the largest eigenvalues are selected (i.e.,  $n_{\text{components}} = 3$ ) to form the principal components. This selection balances reducing dimensionality while retaining significant information. However, to fully explore the impact on classification performance, we also analyze the effects of selecting different numbers of principal components  $p$ .

The transformed dataset  $\text{PCA}_{\text{FD}}$  is given by:

$$\text{PCA}_{\text{FD}} = \text{FD} \cdot \mathbf{W} \quad (11)$$

where  $\mathbf{W}$  is the matrix of the selected eigenvectors. By adjusting the number of retained components  $p$ , we create a compact representation of the data that highlights the most important variations in the gaze trajectories. Specifically, the transformed feature vector  $\text{PCA}_{\text{FD}}$  is:

$$\text{PCA}_{\text{FD}} = [\text{PCA}_{\text{FD}1}, \text{PCA}_{\text{FD}2}, \dots, \text{PCA}_{\text{FD}p}] \quad (12)$$

These components represent the most significant variations in the gaze trajectories captured by the frequency domain features, capturing both the overall trajectory shape and the finer details of oscillatory behaviors.

3) **Time-Lagged Correlation Analysis:** In the time domain, we analyze the horizontal ( $x$ ) and vertical ( $y$ ) coordinates of fixations into a complex signal (as shown in Equation 13) to effectively capture their synchrony, phase differences, and interactions. We extracted the features of time domain using autocorrelation [60]. Autocorrelation quantifies how a signal relates to a time-shifted version of itself, revealing repetitive patterns and periodicity in fixation trajectories. By computing autocorrelation in the complex domain, we can characterize both magnitude and directional consistency of movements over time, which would be difficult to achieve by treating  $x$  and  $y$  separately.

$$z(n) = x(n) + iy(n) \quad (13)$$

where  $i$  is the imaginary unit ( $i^2 = -1$ ).

The autocorrelation of the complex signal  $z(n)$  is defined as:

$$R_z(t) = \sum_{n=0}^{N-t-1} z(n) \cdot \overline{z(n+t)} \quad (14)$$

where  $\overline{z}$  denotes the complex conjugate of  $z$ . This captures joint variations in  $x$  and  $y$  over time while preserving phase relationships.

To obtain a scale-invariant measure of periodicity, we normalize by the zero-lag value  $R_z(0)$ :

$$r_z(t) = \frac{R_z(t)}{R_z(0)} \quad (15)$$

The normalized autocorrelation coefficients  $r_z(t)$  quantify repetitive movement patterns while ensuring comparability across different signal magnitudes.

To determine the optimal feature set, we evaluate  $r_z(t)$  over a range of lag values ( $t$ ) and select those maximizing classification accuracy across four detection types:

$$\text{Autocorr} = [r_z(1), r_z(2), \dots, r_z(T)] \quad (16)$$

A complex signal inherently comprises a real and an imaginary part, each capturing different aspects of trajectory dynamics. To fully characterize these dynamics, we extract and analyze these components separately:

$$\begin{aligned} \text{Re}(r_z(t)) = & \frac{\sum_{n=0}^{N-t-1} \text{Re}(z(n)) \cdot \text{Re}(z(n+t))}{\sum_{n=0}^{N-1} |z(n)|^2} \\ & - \frac{\sum_{n=0}^{N-t-1} \text{Im}(z(n)) \cdot \text{Im}(z(n+t))}{\sum_{n=0}^{N-1} |z(n)|^2} \end{aligned} \quad (17)$$

$$\begin{aligned} \text{Im}(r_z(t)) = & \frac{\sum_{n=0}^{N-t-1} \text{Re}(z(n)) \cdot \text{Im}(z(n+t))}{\sum_{n=0}^{N-1} |z(n)|^2} \\ & + \frac{\sum_{n=0}^{N-t-1} \text{Im}(z(n)) \cdot \text{Re}(z(n+t))}{\sum_{n=0}^{N-1} |z(n)|^2} \end{aligned} \quad (18)$$

Equations (17) and (18) decompose  $r_z(t)$  into real and imaginary parts to capture different movement dynamics. The real part reflects synchrony between  $x$  and  $y$  movements at lag  $t$ , such as periodic scanning, while the imaginary part captures asymmetric interactions, like out-of-phase shifts, leveraging the phase information inherent in complex signals.

From the resulting real and imaginary parts of  $r_z(t)$ , we extract four features : their respective variances and maximum values.

$$\text{Autocorr}_{\text{RealMax}} = \max(\text{Re}(r_z(t))) \quad (19)$$

$$\text{Autocorr}_{\text{ImagMax}} = \max(\text{Im}(r_z(t))) \quad (20)$$

$$\text{Autocorr}_{\text{RealVar}} = \text{Var}(\text{Re}(r_z(t))) \quad (21)$$

$$\text{Autocorr}_{\text{ImagVar}} = \text{Var}(\text{Im}(r_z(t))) \quad (22)$$

The maximum values capture nonlinear peak correlations, while the variances summarize overall periodicity and interaction patterns, enhancing the representation of complex trajectory behaviors. These features provide insights into the visual tracking mechanisms underlying fixation dynamics.

### E. Classification models

In this study, ten machine learning models were systematically evaluated for their effectiveness in detection type recognition. The basic classifiers included support vector machine (SVM) [61], known for handling high-dimensional spaces; k-Nearest Neighbors (kNN) [17], valued for its simplicity and pattern recognition; and multilayer perceptron (MLP), a neural network adept at capturing non-linear relationships [62]. Advanced ensemble methods, including random forest [31], eXtreme Gradient Boosting (XGBoost), and gradient boosting decision tree (GBDT) and Light Gradient Boosting Machine (LightGBM), were also tested for their robustness and efficiency in complex datasets [63]. Additionally, three state-of-the-art deep learning models were incorporated: Transformer-MLP, which uses multi-head self-attention to encode fixation sequences and an MLP for classification, capturing long-range dependencies [64]; LSTM-Attn, a bidirectional LSTM with attention to model structured dependencies and highlight key fixation patterns [36]; and CNN-1D, employing 1D convolutional layers and max pooling to extract local gaze dynamics for hierarchical feature classification [65]. To ensure robust model evaluation and mitigate overfitting, we employed 5-fold cross-validation with an independent test set (20%), for generalization across all models. For deep learning models, we applied L2 regularization (weight decay = 1e-5) and Dropout (0.3) to reduce parameter overfitting, alongside an adaptive learning rate schedule (ReduceLROnPlateau) to optimize training. Additionally, the Synthetic Minority Over-sampling Technique (SMOTE) and Min-Max normalization were used to balance the dataset and standardize feature scales, minimizing bias and enhancing robustness [66]. The models were rigorously evaluated for classifying the four detection types, using accuracy for overall performance and recall for detecting critical failures, minimizing misses of subtle errors [67].

Furthermore, our proposed feature set was utilized to enhance the classification of four visual detection types, compared against twelve traditional eye-tracking features, matched to our proposed set's size for consistency. Traditional features include average fixation duration, standard deviation of fixation duration, fixation frequency, fixation count, saccade count,

saccade frequency, blink rate mean, mean pupil size (left and right), fixation-saccade ratio, standard deviation of saccade magnitude, and maximum saccade magnitude, as extensively documented in previous research on eye movement analysis [21], [34], [57], [68].

## IV. RESULTS

This section evaluates the effectiveness of our proposed hierarchical detection failure recognition framework through phase-specific state-of-the-art comparisons, as the framework addresses the spatiotemporal uncertainty of alerts and considers peripheral vision, making direct comparisons with pre-defined AOI-based frameworks challenging due to their inability to handle such randomness and visual factors. Detailed results follow.

### A. Adaptive symbolic alert detection performance

In this study, we implemented the Adaptive Symbolic Alert Detection method to evaluate its effectiveness in accurately detecting and annotating alerts within an ATC simulation. To assess its performance, we collected 442 valid samples with effective eye-tracking and response data, as detailed in Appendix A. Our proposed method successfully detected and annotated alert regions in 421 of these samples, achieving an accuracy of 95.24%. This metric demonstrates the robustness of our method in accurately identifying alert regions within ATC radar-based alert scenarios, specifically ensuring the complete capture of aircraft icons and corresponding flight information. As a result, the 421 successfully annotated samples were selected for subsequent analysis.

Moreover, to compare our approach with state-of-the-art methods, we evaluated our Adaptive Symbolic Alert Detection method against the Segment Anything Model (SAM) [69] on the same dataset of 442 images. SAM is a general-purpose, transformer-based model designed for segmenting objects in diverse scenarios with high adaptability, which makes it a suitable benchmark for evaluating our adaptive approach. The comparison results show that our method showed better accuracy and efficiency. Specifically, our method correctly identified 421 segmented alert regions, compared to SAM's 405, resulting in a modest difference (95.24% vs. 91.63%). In terms of efficiency, our method achieved a total execution time of 2.883 seconds, averaging 0.007 seconds per image, while SAM required 14086.439 seconds total, averaging 31.870 seconds per image, highlighting a substantial processing time advantage for real-time applications. The SAM method's lower performance in domain-specific ATC applications stems from its general-purpose, transformer-based architecture. While it demonstrates high adaptability and robustness in diverse scenarios, its reliance on complex computations and a rigorous processing pipeline inherently compromises operational efficiency compared to our task-tailored methods [70]. For qualitative comparison, we present 12 sample images in Fig. 8, with 6 images each from our method and SAM, showcasing both consistent detections and notable differences.

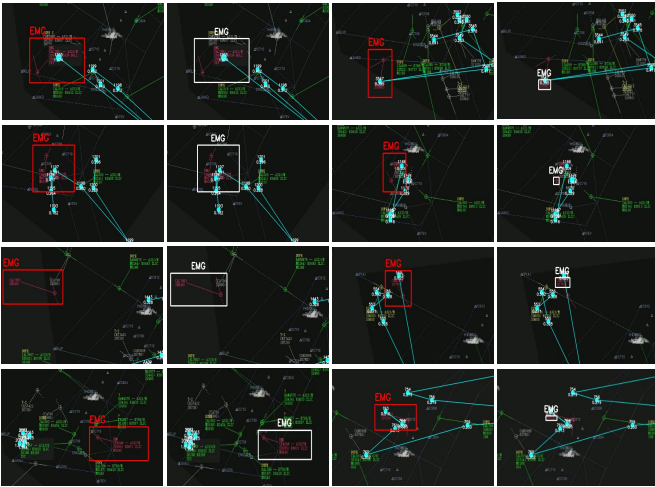


Fig. 8: Qualitative comparison of the proposed Adaptive Symbolic Alert Detection method (red annotations) and SAM method (white annotations) on ATC alert-region detection, showing sample cases with both consistent and divergent outcomes. The proposed method achieves higher accuracy and efficiency, capturing complete aircraft icons and flight information more reliably than SAM.

### B. Visual attention engagement assessment

An LLM-based visual processing approach was used to classify visual attention engagement on annotated alerts into Captured and Non-captured alerts. A sample output illustrating this process is provided in Appendix B. The 421 successfully annotated alert samples from the previous section were categorized into four expert-derived ground truth labels: Effective Detection (291 samples), Peripheral Vision Neglect (52 samples), LBFTS Error (21 samples), and Pre-attentive Detection (57 samples). These labels serve as the reference standard for evaluating the LLM’s accuracy in processing visual attention patterns. As shown in Fig. 9, the LLM’s evaluations were divided into Captured and Non-captured alerts, further distinguishing correct and incorrect assessments. Among the 312 Captured alert cases, GPT-4o correctly identified 231 instances (74.04%), while 81 cases were misclassified. For the 109 Non-captured alert instances, GPT-4V correctly assessed 61 cases (55.96%) but misclassified 48 cases.

The results indicate that GPT-4V performs more effectively in scenarios where alerts were captured, likely due to clearer fixation overlaps with alert regions, reflecting its strength in processing straightforward visual inputs. However, its performance declines in Non-captured scenarios, which are inherently more challenging due to the complexity and ambiguity of peripheral or erratic eye movement trajectories, sample imbalance (312 Captured vs. 109 Non-captured alerts), and limitations in LLM’ pretraining on general-purpose images, underrepresenting dynamic ATC contexts. This disparity suggests that while GPT-4V demonstrates potential for processing captured alerts, its accuracy in ambiguous, Non-captured cases, such as Peripheral Vision Neglect, requires refinement. Overall, these findings underscore the potential of

integrating LLM-based approaches with our advanced multi-domain detection algorithms, enhancing alert detection systems in complex operational environments like ATC.

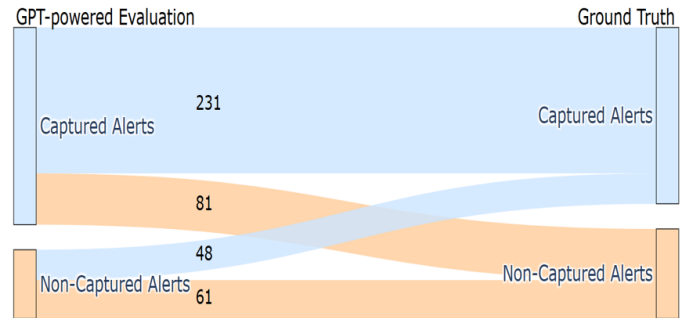


Fig. 9: Performance of LLM (GPT-4V) evaluation of visual attention to alerts, compared against expert-derived ground truth, showing correct and incorrect classifications.

### C. Domain-specific performance in distinguishing visual detection failures

1) **Spatial analysis:** The analysis of multi-dimensional spatial features across four detection types provides crucial insights into how different visual detection types are characterized by specific eye movement patterns. To ensure a more reliable and representative comparison, the Mann-Whitney U test was conducted based on the average values of each spatial feature for each participant, rather than using the overall sample size. This participant-level analysis reduces the influence of intra-individual variability, offering a more accurate assessment of between-group differences. The Mann-Whitney U test, a non-parametric method, was chosen for its robustness in handling non-normal data and effectiveness in comparing independent groups [71]. The four spatial features examined reveal distinct patterns among the detection types, particularly highlighting statistically significant differences, as shown in Fig. 10. In analyzing visual scanning behavior across different detection types, several key patterns emerge across the subfigures. In subfigures (a) and (b), the analysis of gaze stability and fixation dispersion reveals consistent patterns between “Peripheral Vision Neglect” and “Effective Detection.” “Peripheral Vision Neglect” exhibits a significantly more concentrated distribution of gaze points within fixations ( $p = 3.939e-03$ ) and lower fixation dispersion ( $p = 8.547e-04$ ) compared to “Effective Detection.” These findings indicate a more concentrated and stable visual focus during “Peripheral Vision Neglect,” suggesting that participants’ intense focus on central stimuli may limit their ability to effectively track and respond to peripheral cues, leading to the neglect of peripheral stimuli. In subfigure (c), the number of fixation clusters shows significant differences between “LBFTS” errors and both “Effective Detection” ( $p = 3.800e-03$ ) and “Pre-attentive Detection” ( $p = 3.843e-02$ ), with “LBFTS” errors having fewer clusters. This suggests that during LBFTS errors, although the alert is perceived, the limited fixation clusters indicate insufficient cognitive resource distribution for effective processing and interpretation. This cognitive disconnect leads

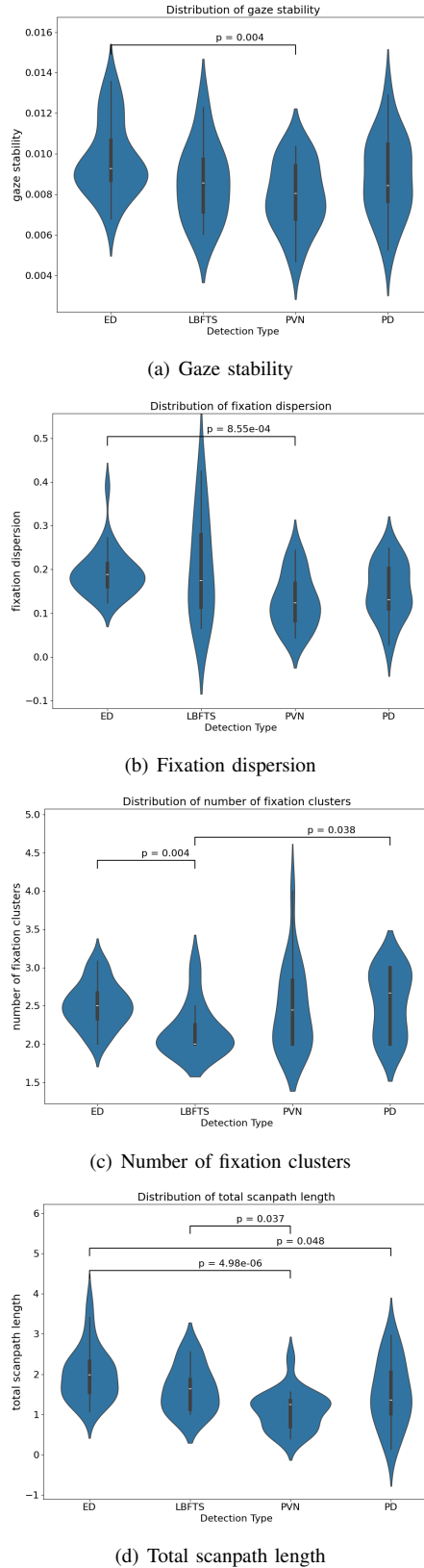


Fig. 10: Spatial feature analysis across four detection types: ED (Effective Detection), PVN (Peripheral Vision Neglect), LBFTS (Look-but-fail-to-see Error), and PD (Pre-attentive Detection), with significant differences via Mann-Whitney U test. (Note: Horizontal lines indicate significant differences between pairs, with p-values shown above each line.)

to an inadequate response despite initial detection. In subfigure (d), the analysis of total scanpath length reveals significant differences between "Peripheral Vision Neglect" and both "LBFTS" errors ( $p = 3.683e-02$ ) and "Effective Detection" ( $p = 4.981e-06$ ). "Peripheral Vision Neglect" shows the shortest scanpath length overall, indicating that participants' eye movements are most constrained during this condition. This suggests a highly focused visual attention on specific tasks, leading to the neglect of peripheral alerts. In addition, the significant difference between "Pre-attentive Detection" and "Effective Detection" ( $p = 4.753e-02$ ), suggests that when participants perceive an alert peripherally and respond without direct fixation, their visual search is more efficient, requiring less extensive eye movement. This efficiency likely stems from quickly detecting and processing the alert within peripheral vision, reducing the need for broader scanning typical in "Effective Detection".

Collectively, these findings across the subfigures underscore the distinct visual scanning behaviors and the importance of spatial features associated with different detection types, providing valuable insights into the underlying mechanisms that contribute to detection successes and failures.

2) **Frequency Domain Analysis:** This section evaluates the classification performance of frequency-domain features in distinguishing peripheral vision-aware detection failures in ATC. By isolating Fourier Descriptors (FDs) and PCA components, we assess their contributions to identifying four visual detection types. Random Forest, the best-performing model among ten classifiers, achieved the highest test precision across all configurations.

Fig. 11 illustrates the classification accuracies for FD ( $k$ ) ranging from 3 to 10 and PCA components ( $p$ ) from 2 to 4, selected to balance capture sufficient frequency detail while avoiding noise in Discrete Fourier Transforms, with PCA limited to prevent redundancy. In general, accuracy peaks at 82% with  $k = 5$  and  $p = 4$ , suggesting that this combination optimally balances the richness of Fourier descriptors and the reduction in dimensionality provided by PCA. Interestingly, the middle values of  $k$  (e.g. 6) with lower  $p$  (e.g., 2) tend to result in lower precision (61%), which could be due to the descriptors in these values not providing enough variation or being less representative of the shape characteristics needed for optimal classification. In contrast, lower values of  $k$  (e.g., 3) also lead to reduced accuracy (64%), likely due to insufficient descriptor complexity. The trend suggests that increasing  $k$  improves accuracy up to a point, but the model benefits the most with the  $p = 4$  components for the reduction of dimensionality, as higher  $k$  without appropriate adjustments  $p$  does not consistently improve performance. These findings highlight the importance of carefully tuning FD and PCA parameters to optimize frequency feature extraction for complex classification tasks, revealing their role in capturing nuanced gaze dynamics critical for ATC detection.

To further explore the frequency domain's role across detection types, we examined the per-class performance of the optimal configuration ( $k=5, p=4$ ) using RF, as shown in Fig. 12. The normalized confusion matrix illustrates the classification performance, with the diagonal values representing the

proportion of correctly classified samples for each detection type: Effective Detection (0.72), Peripheral Vision Neglect (0.88), LBFTS Error (0.85), and Pre-attentive Detection (0.84). Additionally, the recall scores were calculated for each class: Effective Detection (0.60), Peripheral Vision Neglect (0.88), LBFTS Error (0.93), and Pre-attentive Detection (0.81). These results demonstrate that frequency features excel at detecting Peripheral Vision Neglect and LBFTS Errors, likely due to their sensitivity to rapid, oscillatory saccades and constrained fixations, respectively, as supported by prior research on saccadic dynamics [56]. However, the lower performance for Effective Detection suggests integrating temporal or spatial features to enhance its detection. Overall, the findings highlight the frequency domain's ability to model rhythmic gaze disruptions, enhancing ATC safety by identifying subtle cognitive lapses and attentional constraints with distinct dynamic patterns.

Random Forest Classification Accuracy across Different FD and PCA Combinations

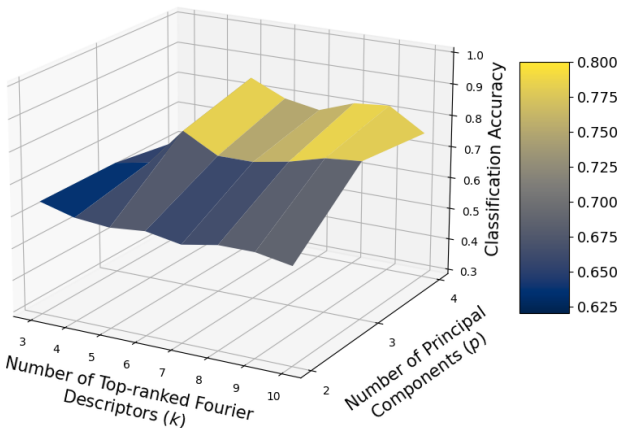


Fig. 11: Random forest classification accuracy for varying Fourier Descriptors ( $k=3-10$ ) and PCA components ( $p=2-4$ ), peaking at 81% with  $k=5$  and  $p=4$  for four visual detection types.

3) **Time-Lagged Correlation Analysis:** This section examines the classification performance of time-lagged autocorrelation features, evaluating their role in distinguishing four detection failures in ATC. By isolating features capturing maximum autocorrelation values and their variances of the real and imaginary parts across lags 2–10, we clarify their contribution to recognizing proposed four visual detection type. Random Forest, the highest-performing model among ten classifiers, achieved the best test accuracy across all configurations. Fig. 13 presents a radar plot of test accuracies for lags ranging from 2 to 10, selected based on typical fixation counts (13–22 per sample) to ensure meaningful temporal patterns without overcomplicating the model. Results show accuracy peaks at 81% for lag=9, with a range from 72% (lag=2) to 80% (lag=8). This trend indicates that longer lags, particularly lag=9, capture sustained attention patterns critical

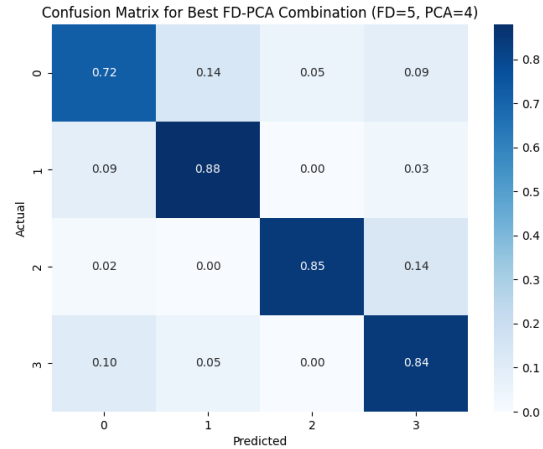


Fig. 12: Confusion matrix for the random forest model using four frequency-domain features, showing performance across four detection types (0: Effective Detection, 1: Peripheral Vision Neglect, 2: LBFTS Error, 3: Pre-attentive Detection )

for ATC detection, revealing their role in modeling temporally distinct gaze behaviors.

To evaluate the performance of time domain features across detection types, we analyzed the per-class performance of the optimal configuration (lag=9) using Random Forest, as shown in Fig. 14. The normalized confusion matrix reveals classification performance, with diagonal values indicating correct classification rates: Effective Detection (0.59), Peripheral Vision Neglect (0.86), LBFTS Error (0.92), and Pre-attentive Detection (0.84). The model excels in classifying LBFTS Error and Peripheral Vision Neglect, reflecting time-lagged features' ability to capture distinct temporal dynamics, such as centrally focused fixations in LBFTS and regular, concentrated patterns in PVN, critical for identifying these failures. However, the lower performance for Effective Detection suggests integrating frequency or spatial features to enhance its detection.

To quantify these temporal dynamics, we analyzed the average maximum autocorrelation values at lag=9. Peripheral Vision Neglect and Pre-attentive Detection exhibit the highest real part maximum values (0.86 each), indicating strong synchronization in fixation coordinates, with low imaginary part values (0.03 and 0.02) suggesting minimal phase differences. LBFTS Error (real 0.83, imag 0.03) and Effective Detection (real 0.82, imag 0.05) show slightly lower synchronization. Despite Peripheral Vision Neglect's strong synchronization, LBFTS Error's superior performance suggests time-lagged features better capture its centrally focused fixation patterns, possibly due to distinct temporal consistency in gaze concentration. For qualitative comparison, Fig. 15 displays sample trajectories (top-ranked by autocorrelation values at time-lag=9), revealing prolonged, inconsistent temporal shifts in Effective Detection, rhythmic and sustained fixation cycles in PVN, stable and prolonged central focus in LBFTS, and variable temporal patterns with intermittent attention shifts

in Pre-attentive Detection, validating the distinct time-lagged behaviors across categories.

Random Forest Classification Accuracy across Different Time Lag ( $t$ )

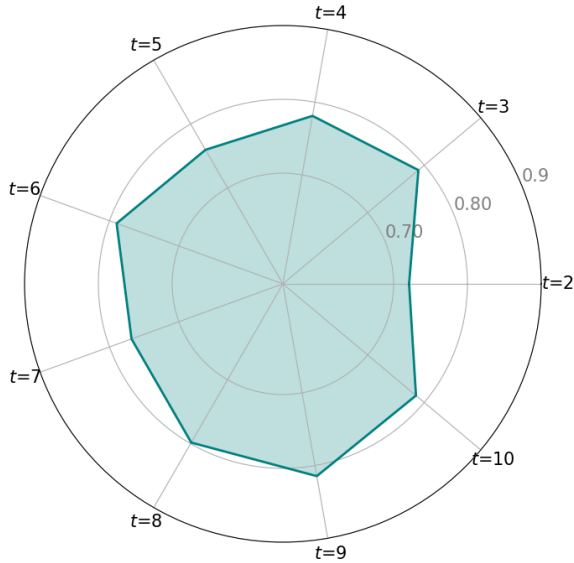


Fig. 13: Random forest classification accuracy across time lags ( $t=2-10$ ) in autocorrelation analysis, with peak performance of 81% at  $t=9$  for four visual detection types.

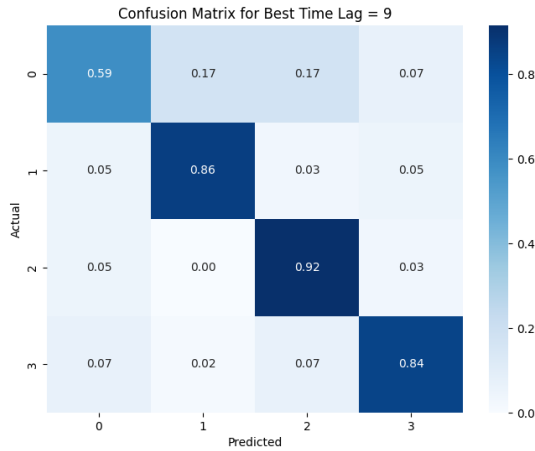


Fig. 14: Confusion matrix for the random forest model using four time-domain features, showing performance across four detection types (0: Effective Detection, 1: Peripheral Vision Neglect, 2: LBFTS Error, 3: Pre-attentive Detection)

#### D. Integrated domain performance in visual detection failure recognition

This study evaluates the classification performance of visual detection failure recognition, focusing on four distinct detection types, by comparing the proposed fixation

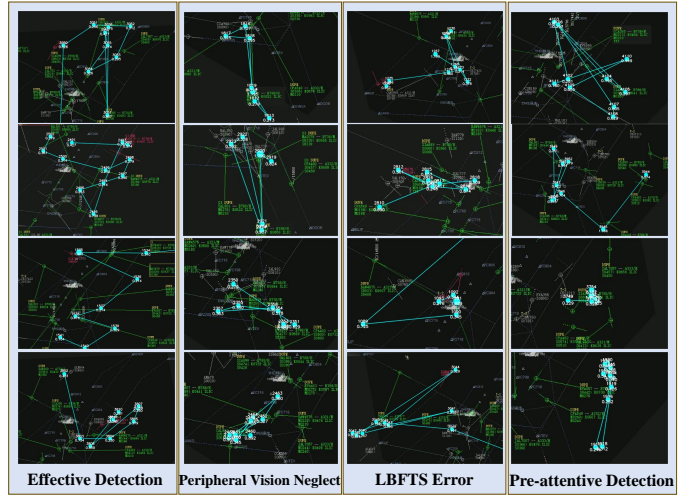


Fig. 15: Qualitative comparison of temporal characteristics across four detection types using samples with top-ranked maximum autocorrelation samples.

coordinate-sensitive multi-domain feature set with several alternatives. The proposed feature set integrates 12 features across three domains—four spatial, four frequency (FD=5, PCA=3, the optimal configuration identified in Section IV-C2), and four temporal (lag=9, the optimal value identified in Section IV-C3)—to capture a comprehensive representation of gaze behavior. For comparison, we include a traditional feature set (also containing 12 features to ensure fairness), alongside domain-specific subsets: single-domain sets (spatial-only, frequency-only, temporal-only, each with four features) and dual-domain combinations (spatial-frequency, spatial-temporal, frequency-temporal).

Table I presents the classification accuracy across ten classifiers, highlighting the robustness of the proposed multi-domain feature set through an ablation study. Firstly, the proposed set consistently outperforms the traditional feature set across all models, achieving the highest accuracy of 93.13% with Transformer-MLP, compared to 82.83% for the traditional set—a 10.30% improvement. LightGBM ranks second at 92.27% (vs. 83.69% for traditional features), demonstrating the proposed set's superior generalization across classifiers. Ablation analysis further reveals the contribution of individual and paired domains. Single-domain subsets perform significantly worse, with spatial features-only yielding the lowest accuracy (peak at 76.82%), followed by temporal features-only (peak at 81.00%) and frequency features-only (peak at 80.69%), underscoring the limitations of relying on isolated domains. Dual-domain combinations, however, show marked improvements over single domains: frequency-temporal features achieve 90.99% with LightGBM, and spatial-frequency features reach 90.30% with Transformer-MLP. Notably, spatial-temporal features underperform (peak at 83.26%), suggesting potential challenges in model convergence for this combination. These results confirm the necessity of integrating all three domains, as the proposed multi-domain set consistently surpasses both single- and dual-domain sub-

TABLE I: Accuracy (%) of traditional and the proposed fixation coordinate-sensitive multi-domain feature sets for four-category detection-type recognition across ten classifiers via ablation analysis of the proposed set with single- and dual-domain tests.

Feature Set	SVM	kNN	MLP	RF	XGBoost	GBDT	LightGBM	Transformer-MLP	LSTM-Attn	CNN-1D
Traditional features set	66.09	79.40	78.11	82.40	82.40	76.82	<b>83.69</b>	82.83	79.40	64.38
The proposed feature set	79.83	81.55	85.84	90.13	88.41	82.83	92.27	<b>93.13</b>	88.84	74.68
Spatial-frequency features	71.24	80.26	77.68	85.84	84.55	76.39	87.12	<b>90.30</b>	69.50	86.27
Spatial-temporal features	64.38	76.82	75.11	<b>83.26</b>	81.12	75.11	<b>83.26</b>	82.00	67.81	79.40
Frequency-temporal features	69.53	80.26	79.40	87.55	87.55	81.97	<b>90.99</b>	89.00	68.00	80.26
Spatial features-only	54.08	72.10	54.94	<b>76.82</b>	69.96	63.52	71.24	71.24	54.94	51.07
Frequency features-only	62.23	71.67	60.09	<b>80.69</b>	77.25	70.82	80.26	76.82	61.37	66.09
Temporal features-only	57.51	69.10	57.51	<b>81.00</b>	79.40	71.24	78.97	71.67	57.51	60.94

sets, highlighting the critical role of domain interplay in enhancing detection-type recognition accuracy.

## V. DISCUSSIONS

This study establishes a novel peripheral vision-aware visual detection classification and validates it by introducing a Hierarchical Detection Failure Recognition Framework, demonstrating the distinct differences between various detection types. We first conducted an ATC simulation experiment to provide empirical data for the framework analysis. The first phase of the framework introduces the Adaptive Symbolic Alert Detection method, which identifies alert regions with spatiotemporal uncertainty on the ATC radar interface. This method enables robust identification of alert information on radar interfaces, achieving a 95.24% precision rate. Our study, aligned with prior research isolating critical events to establish a foundational understanding of psychophysiological patterns [43], [44], currently focuses on single-alert scenarios. Should overlapping alert scenarios be further explored, our method effectively identifies the overall alert regions; however, future research should focus on delineating individual alert boundaries within overlaps and incorporating semantic information and alert prioritization to enhance its applicability in complex ATC environments.

In the second phase, GPT-4V is employed to evaluate visual attention to alerts, facilitating the classification of samples into four proposed detection types. Our findings reveal that the LLM’s evaluation of detected alerts shows a clear advantage in scenarios where alerts were captured accurately, with a correct evaluation rate of 74.04%. This outcome aligns with findings in similar studies, such as the research of [55], where LLMs were integrated with EEG and eye-gaze data, achieving a classification accuracy of over 60%. These studies collectively highlight the potential of LLMs in tasks requiring semantic and contextual understanding, especially when the visual or linguistic data is straightforward and less ambiguous. However, the LLM’s performance demonstrates a need for further refinement in more challenging situations, such as the evaluation of Non-captured alerts where the correct rate dropped to 55.96%.

A potential explanation of the disparity in performance between Captured and Non-captured alert evaluations can be attributed to three key factors. The complexity and ambiguity of Non-captured scenarios, where eye movement trajectories often exhibit peripheral or erratic patterns, challenge GPT-4V’s visual capabilities, as these patterns reflect peripheral

vision or cognitive failures (e.g., attention dispersion, lack of understanding). Second, sample imbalance—312 Captured versus 109 Non-captured alerts—biases LLM toward captured scenarios, consistent with class imbalance theory [72]), which shows models prioritize majority classes, and AI bias studies [73], which highlight reduced sensitivity to rare events. Third, GPT-4V’s pretraining on general-purpose images and texts underrepresents dynamic ATC contexts, limiting its ability to capture subtle differences in peripheral vision or unstable trajectories [74]. To address the challenges, we propose building on the results of this study as a baseline to explore different LLM models with prompt designs, such as chain-of-thought (CoT) prompting for reasoning about gaze patterns or instruction-based prompting with task-specific guidelines for peripheral vision dynamics, and fine-tuning LLMs with ATC-specific datasets. This approach aligns with recent advances in LLM optimization [55] and aims to enhance the model’s ability to capture nuanced eye movement patterns, ultimately improving performance in ATC alert detection tasks.

Additionally, the framework introduces a novel fixation coordinate-sensitive multi-domain feature set, achieving 93.13% accuracy in peripheral vision-aware detection, surpassing traditional eye-tracking metrics (83.69%) and both single- and dual-domain features. Compared to prior research on four visual attentional type classification [56], which achieved only 79% accuracy, our results highlight the significant advantage of fixation coordinate-sensitive multi-domain feature integration in capturing nuanced visual tracking dynamics. The following discussion will examine how the four detection types are differentiated through the employed domain-specific analysis, shedding light on the underlying mechanisms and their potential implications for improving detection accuracy.

In this study, spatial feature analysis offers critical insights into how visual attention is distributed across detection successes and failures. Previous studies, such as [35], have suggested that efficient eye movement patterns are typically characterized by broader and more extensive tracking trajectories, indicating effective visual processing. Consistent with these findings, our study found that LBFTS errors are associated with significantly fewer fixation clusters compared to “Effective Detection” and “Pre-attentive Detection”. Additionally, the significantly higher gaze stability and lower fixation dispersion observed in “Peripheral Vision Neglect” compared to “Effective Detection” align with the challenges in maintaining consistent attention to peripheral stimuli, as highlighted by [75]. Thus, the two detection failures result from difficulties

in processing peripheral information, reflecting a more focused but less exploratory search strategy. Furthermore, the two detection failures are significantly different, as reflected by "Peripheral Vision Neglect" showing significantly shorter scanpath lengths than "LBFTS," indicating over-focus on specific tasks and neglect of peripheral alerts. Besides, the two successful detection types also differ significantly, with "Pre-attentive Detection" exhibiting shorter lengths than "Effective Detection," indicating a more efficient, less extensive search during peripheral detection without direct fixation.

The introduction of frequency domain analysis has proven to be a powerful tool for distinguishing between different detection types by analyzing how varying the number of FDs and PCA components impacts recognition accuracy. Previous studies, such as the examination of saccadic eye movements in [58], have highlighted the importance of FDs in capturing dynamic visual patterns. Similarly, [76] demonstrated that incorporating frequency domain features can significantly enhance temporal information processing in emotion recognition tasks. In our study, we found that increasing the FD generally improves accuracy, with a peak observed at  $FD=5$  and  $PCA=4$ , achieving an accuracy of 81%. Furthermore, the strong performance in detecting LBFTS Errors and Peripheral Vision Neglect underscores the frequency domain's sensitivity to rhythmic gaze disruptions, such as oscillatory saccades and constrained fixations, offering a reliable approach to identify critical cognitive lapses in high-stakes ATC settings. However, its limitation in capturing dispersed gaze patterns, as observed in Effective Detection, underscores the need for multi-domain frameworks integrating time, frequency, and spatial features. Additionally, while our noise mitigation strategies minimize distortion, advanced denoising techniques like wavelet transforms or adaptive filtering could further suppress high-frequency jitters and stabilize visual trajectory patterns, addressing residual noise concerns [77].

Building on the insights from the time-lagged correlation analysis, this discussion deepens the understanding of temporal dynamics across detection types, with the highest accuracy 81% at  $lag=9$  underscoring the importance of capturing sustained attention patterns essential for identifying cognitive lapses such as LBFTS Errors and Peripheral Vision Neglect. This aligns with [12], which emphasizes the critical role of temporal dependencies in visual attention mechanisms, a finding reinforced by [78], who linked LBFTS failures to ineffective temporal processing. The robust performance in detecting centrally focused fixations and rhythmic gaze disruptions further validates a strong framework for pinpointing high-risk attentional failures; however, the difficulty in capturing dispersed patterns in Effective Detection highlights the necessity for adaptive multi-domain approaches. These observations inspire future research into dynamic lag optimization, real-time integration of temporal features with spatial and frequency domains, and personalized gaze modeling to accommodate diverse attentional dynamics, thereby enhancing the resilience of automated ATC systems by adapting to individual operator variability and minimizing detection errors.

The proposed hierarchical peripheral vision-aware framework enhances ATC as a critical transportation system by

dynamically assessing operator attentional states to improve alert systems, effectively differentiating meaningful alert engagement from mere gaze direction—a distinction vital in ATC where missed warnings stem not from lacking visual contact but from attentional prioritization failures [56]. Building on this, the framework advances beyond immediate detection by fostering an intelligent, human-centered ATC support system that identifies detection failures and informs adaptive intervention strategies. By analyzing visual attention fluctuations in dynamic air traffic scenarios, it enables personalized alert optimization, adjusting warning intensity and format to individual controller engagement patterns. This shift from reactive monitoring to proactive enhancement strengthens ATC resilience against attentional lapses, preserving situational awareness and decision-making integrity in high-stakes settings. Future research should focus on practical deployment strategies and real-time implementation feasibility within ATC centers.

## VI. CONCLUSIONS

This study introduced a novel classification framework for visual detection results incorporating peripheral vision awareness and human attentional states, highlighting a broader and more structured range of visual detection failures. Furthermore, a hierarchical framework has been developed to effectively recognize the proposed detection failures in ATC-specific environments. First, our Adaptive Symbolic Alert Detection method successfully addressed the spatiotemporal uncertainties of alert appearances on ATC radar interfaces, achieving a precision rate of 95.24%. Additionally, the GPT-4V-powered LLMs showed significant potential in enhancing the verification of controllers' visual attentional engagement with alerts, extending the capabilities beyond what traditional eye-tracking methods alone could achieve. Finally, the proposed fixation coordinate-sensitive multi-domain feature set markedly enhanced detection-type recognition, with accuracy rising from 83.69% (traditional features) to 93.13% with Transformer-MLP, outperforming single-domain representations (spatial: 76.82%, temporal: 81.00% and frequency: 80.69%), dual-domain combinations (ranging from 83.26% to 90.99%) and other models like LightGBM (92.27%) and Random Forest (90.13%), demonstrating the synergistic effectiveness of multi-domain feature integration.

However, this study has limitations that warrant further investigation. First, this study focused on visual-only alerts to explore peripheral vision tracking in ATC scenarios, where alerts may use visual cues alone or combine visual and auditory signals based on importance. Future research should include multi-modal alerts to better understand detection failures in settings with both visual and auditory components, potentially integrating spoken instructions as explored by [79] to enhance detection failure recognition. While we focused on single radar-based alerts for foundation understanding of visual tracking patterns, future work should explore overlapping alerts to evaluate the framework's robustness in complex ATC scenarios. In addition, EEG data will be integrated—deferred here to prioritize eye-tracking's robust performance and maintain methodological focus—to explore gaze-neural correla-

tions. Lastly, the LLM's evaluation is computationally intensive, necessitating optimized deployment strategies such as lightweight model compression or specialized inference acceleration for real-time ATC applicability, and the zero-shot prompt's accuracy limitations suggest fine-tuning to better align with specific alert detection tasks.

## APPENDIX

### A. Sample distribution across participants

From an initial pool of 38 participants, 7 were excluded due to invalid data: 3 for eye-tracking calibration failure, 3 for Euroscope platform timing issues preventing alert timing confirmation, and 1 for missing platform log data. Of the 494 total samples collected, 442 (89%) were valid and 52 (11%) were invalid, distributed across the remaining 31 participants as shown in Fig. 16. Invalid samples resulted from technical factors: (1) network delays in the Euroscope platform, causing alerts to exceed the intended 5-second duration (e.g., 6 seconds), and (2) track loss due to head movements beyond the eye tracker's range, leading to gaps in gaze data.

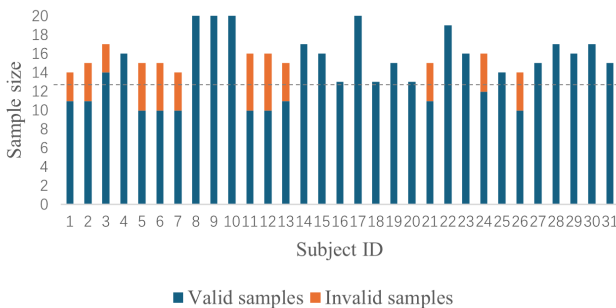


Fig. 16: Distribution of valid and invalid samples per participant. The dashed line represents that participant received at least 13 alerts.

### B. A sample output from GPT-4V Model

Output: The provided image shows a red rectangle in the upper center of the image, marked with "EMG" indicating an alert. The blue points and lines within the image represent the eye movement trajectory of an operator.

Upon close inspection:

1. Identifying red rectangle and blue trajectory:

- The red rectangle is clearly marked at the top center of the image.

- There are several blue points and lines observed in the vicinity of this red rectangle.

2. Assessing the blue points or lines inside the red rectangle:

- There are clearly multiple blue points and lines (eye fixation and saccades) inside the red rectangle.

Based on the criteria provided:

- Since there are blue points and lines inside the red rectangle, this indicates that the subject's attention was captured by the emergency.

Final Classification: Captured alert

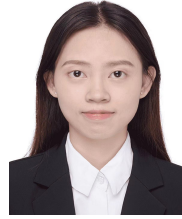
## REFERENCES

- [1] R. Xu and F. Luo, "Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest," *Safety science*, vol. 135, p. 105125, 2021.
- [2] C. Bruder and C. Hasse, "What the eyes reveal: Investigating the detection of automation failures," *Applied ergonomics*, vol. 82, p. 102967, 2020.
- [3] S. T. Shorrock, "Errors of perception in air traffic control," *Safety science*, vol. 45, no. 8, pp. 890–904, 2007.
- [4] J. Liu, Y. Chen, H. Liu, H. Zhang, and Y. Zhang, "From less to more: Progressive generalized zero-shot detection with curriculum learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19 016–19 029, 2022.
- [5] L. Giraudet, J.-P. Imbert, S. Tremblay, and M. Causse, "High rate of inattention deafness in simulated air traffic control tasks," *Procedia Manufacturing*, vol. 3, pp. 5169–5175, 2015.
- [6] K. D. Kennedy, C. L. Stephens, R. A. Williams, and P. C. Schutte, "Automation and inattention blindness in a simulated flight task," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2014, pp. 2058–2062.
- [7] S. Park, C. Y. Park, C. Lee, S. H. Han, S. Yun, and D.-E. Lee, "Exploring inattention blindness in failure of safety risk perception: Focusing on safety knowledge in construction industry," *Safety science*, vol. 145, p. 105518, 2022.
- [8] Z. Li, Z. Li, and F. Li, "Visual attention analytics for individual perception differences and task load-induced inattention blindness," in *International Conference on Human-Computer Interaction*. Springer, 2023, pp. 71–83.
- [9] D. J. Simons and M. S. Jensen, "The effects of individual differences and task difficulty on inattention blindness," *Psychonomic Bulletin & Review*, vol. 16, pp. 398–403, 2009.
- [10] J. Xu, K. Guo, X. Zhang, and P. Z. Sun, "Left gaze bias between lht and rht: a recommendation strategy to mitigate human errors in left-and right-hand driving," *IEEE Transactions on Intelligent Vehicles*, 2023.
- [11] J. Xu, S. H. Park, X. Zhang, and J. Hu, "The improvement of road driving safety guided by visual inattention blindness," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 6, pp. 4972–4981, 2021.
- [12] C.-Y. Tse, Y.-H. Shum, X.-Z. Xiao, and Y. Wang, "Fronto-occipital mismatch responses in pre-attentive detection of visual changes: Implication on a generic brain network underlying mismatch negativity (mmn)," *Neuroimage*, vol. 244, p. 118633, 2021.
- [13] J. M. Wolfe, A. Kosovicheva, and B. Wolfe, "Normal blindness: When we look but fail to see," *Trends in Cognitive Sciences*, vol. 26, no. 9, pp. 809–819, 2022.
- [14] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.
- [15] S. Han, F. Li, C.-H. Lee, T. Wang, and M. A. Diaconeasa, "Mirror the mind of crew: Maritime risk analysis with explicit cognitive processes in a human digital twin," *Advanced Engineering Informatics*, vol. 62, p. 102746, 2024.
- [16] A. Lleras, S. Buetti, and Z. J. Xu, "Incorporating the properties of peripheral vision into theories of visual search," *Nature Reviews Psychology*, vol. 1, no. 10, pp. 590–604, 2022.
- [17] J. Yang, N. Liang, B. J. Pitts, K. Prakah-Asante, R. Curry, and D. Yu, "An eye-fixation related electroencephalography technique for predicting situation awareness: implications for driver state monitoring systems," *Human factors*, vol. 66, no. 8, pp. 2138–2153, 2024.
- [18] L. Zhaoping and P. Dayan, "Pre-attentive visual selection," *Neural Networks*, vol. 19, no. 9, pp. 1437–1439, 2006.
- [19] J. H. Lee and M. J. Ostwald, "The 'visual attractiveness' of architectural facades: measuring visual complexity and attractive strength in architecture," *Architectural science review*, vol. 66, no. 1, pp. 42–52, 2023.
- [20] J.-P. Imbert, H. M. Hodgetts, R. Parise, F. Vachon, F. Dehais, and S. Tremblay, "Attentional costs and failures in air traffic control notifications," *Ergonomics*, vol. 57, no. 12, pp. 1817–1832, 2014.
- [21] M. Lyu, F. Li, C.-H. Lee, and C.-H. Chen, "Valio: Visual attention-based linear temporal logic method for explainable out-of-the-loop identification," *Knowledge-Based Systems*, p. 112086, 2024.
- [22] L. Qin, Y. Shi, Y. He, J. Zhang, X. Zhang, Y. Li, T. Deng, and H. Yan, "Id-yolo: Real-time salient object detection based on the driver's fixation region," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15 898–15 908, 2022.

- [23] V. Onkhar, L. Kumaaravelu, D. Dodou, and J. De Winter, "Towards context-aware road user safety systems: Design explorations using eye-tracking, object detection, and gpt-4v," 2024.
- [24] Y. Zhang, F. Li, and D. Chang, "Vr rehabilitation system evaluator: A fnirs-based and llm-enabled evaluation paradigm for mild cognitive impairment," *Advanced Engineering Informatics*, vol. 62, p. 102734, 2024.
- [25] M. Z. Hasan, J. Chen, J. Wang, M. S. Rahman, A. Joshi, S. Velipasalar, C. Hegde, A. Sharma, and S. Sarkar, "Vision-language models can identify distracted driver behavior from naturalistic videos," *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [26] X. Wang, H. Kim, S. Rahman, K. Mitra, and Z. Miao, "Human-llm collaborative annotation through effective verification of llm labels," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–21.
- [27] F. Li, S. Feng, Y. Yan, C.-H. Lee, and Y. S. Ong, "Virtual co-pilot: Multimodal large language model-enabled quick-access procedures for single pilot operations," in *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 1501–1506.
- [28] S. Shahriar, B. D. Lund, N. R. Mannuru, M. A. Arshad, K. Hayawi, R. V. K. Bevara, A. Mannuru, and L. Batool, "Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency," *Applied Sciences*, vol. 14, no. 17, p. 7782, 2024.
- [29] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Robust visual tracking via hierarchical convolutional features," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2709–2723, 2018.
- [30] F. Li, C.-H. Chen, G. Xu, L. P. Khoo, and Y. Liu, "Proactive mental fatigue detection of traffic control operators using bagged trees and gaze-bin analysis," *Advanced Engineering Informatics*, vol. 42, p. 100987, 2019.
- [31] F. Zhou, X. J. Yang, and J. C. De Winter, "Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2284–2295, 2021.
- [32] F. Li, C.-H. Chen, C.-H. Lee, and S. Feng, "Artificial intelligence-enabled non-intrusive vigilance assessment approach to reducing traffic controller's human errors," *Knowledge-Based Systems*, vol. 239, p. 108047, 2022.
- [33] H. Yu and M. Kwon, "Altered eye movements during reading with simulated central and peripheral visual field defects," *Investigative Ophthalmology & Visual Science*, vol. 64, no. 13, pp. 21–21, 2023.
- [34] Z. Li, R. Li, L. Yuan, J. Cui, and F. Li, "A benchmarking framework for eye-tracking-based vigilance prediction of vessel traffic controllers," *Engineering Applications of Artificial Intelligence*, vol. 129, p. 107660, 2024.
- [35] X. Ma, Y. Liu, R. Clariana, C. Gu, and P. Li, "From eye movements to scanpath networks: A method for studying individual differences in expository text reading," *Behavior research methods*, vol. 55, no. 2, pp. 730–750, 2023.
- [36] H. Qin, H. Zhu, X. Jin, Q. Song, M. A. El-Yacoubi, and X. Gao, "Emmixformer: Mix transformer for eye movement recognition," *arXiv preprint arXiv:2401.04956*, 2024.
- [37] K. Ning, Y. Liu, Y. Su, and K. Jiang, "Point-voxel and bird-eye-view representation aggregation network for single stage 3d object detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3223–3235, 2022.
- [38] Q. Li, K. K. Ng, C. Y. Yiu, X. Yuan, C. K. So, and C. C. Ho, "Securing air transportation safety through identifying pilot's risky vfr flying behaviours: An eeg-based neurophysiological modelling using machine learning algorithms," *Reliability Engineering & System Safety*, vol. 238, p. 109449, 2023.
- [39] Q. Li, K. K. Ng, C. Simon, C. Y. Yiu, and M. Lyu, "Recognising situation awareness associated with different workloads using eeg and eye-tracking features in air traffic control tasks," *Knowledge-Based Systems*, vol. 260, p. 110179, 2023.
- [40] Y. Zhang, Y. Liao, Y. Zhang, and L. Huang, "Emergency braking intention detect system based on k-order propagation number algorithm: a network perspective," *Brain Sciences*, vol. 11, no. 11, p. 1424, 2021.
- [41] W. S. Helton, G. Matthews, and J. S. Warm, "Stress state mediation between environmental variables and performance: The case of noise and vigilance," *Acta psychologica*, vol. 130, no. 3, pp. 204–213, 2009.
- [42] K. G. Tippey, T. Roady, C. Rodriguez-Paras, L. J. Brown, W. G. Rantz, and T. K. Ferris, "General aviation weather alerting: the effectiveness of different visual and tactile display characteristics in supporting weather-related decision making," *The International Journal of Aerospace Psychology*, vol. 27, no. 3-4, pp. 121–136, 2017.
- [43] F. Trapsilawati, X. Qu, C. D. Wickens, and C.-H. Chen, "Human factors assessment of conflict resolution aid reliability and time pressure in future air traffic control," *Ergonomics*, vol. 58, no. 6, pp. 897–908, 2015.
- [44] P. Kearney, W.-C. Li, and J. J. Lin, "The impact of alerting design on air traffic controllers' response to conflict detection and resolution," *International journal of industrial ergonomics*, vol. 56, pp. 51–58, 2016.
- [45] T. C. Farley, R. J. Hansman, K. Amonlirdviman, and M. R. Endsley, "Shared information between pilots and controllers in tactical air traffic control," *Journal of Guidance, Control, and Dynamics*, vol. 23, no. 5, pp. 826–836, 2000.
- [46] K. R. Allendoerfer, F. Friedman-Berg, S. Pai *et al.*, "Human factors analysis of safety alerts in air traffic control," United States. Department of Transportation. Federal Aviation Administration ..., Tech. Rep., 2007.
- [47] T. Jürgensohn, I. Park, T. B. Sheridan, and J. Meyer, "Conflict warnings and the search and detection of collisions in a simulated atc-like task," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 45, no. 4. SAGE Publications Sage CA: Los Angeles, CA, 2001, pp. 458–462.
- [48] K. J. Ruskin, S. Rice, C. Corvin *et al.*, "Atc signaling systems: A review of the literature on alarms, alerts, and warnings," 2020.
- [49] F. Friedman-Berg, K. Allendoerfer, and S. Pai, "Nuisance alerts in operational atc environments: Classification and frequencies," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2008, pp. 104–108.
- [50] A. Mack, "Inattentive blindness: Looking without seeing," *Current directions in psychological science*, vol. 12, no. 5, pp. 180–184, 2003.
- [51] C. Binatha, "Hue saturation value (hsv) color space for content based image retrieval," *International Journal of Engineering Research*, vol. 5, no. 04, 2016.
- [52] G. Yong, K. Jeon, D. Gil, and G. Lee, "Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 11, pp. 1536–1554, 2023.
- [53] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological bulletin*, vol. 124, no. 3, p. 372, 1998.
- [54] L. Chamberlain, "Eye tracking methodology; theory and practice," *Qualitative Market Research: An International Journal*, vol. 10, no. 2, pp. 217–220, 2007.
- [55] Y. Zhang, Q. Li, S. Nahata, T. Jamal, S.-K. Cheng, G. Cauwenberghs, and T.-P. Jung, "Integrating large language model, eeg, and eye-tracking for word-level neural state classification in reading comprehension," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.
- [56] Z. Li, F. Li, and M. Lyu, "Tracking the unseen and unaware: Deciphering controllers' detection failures to warnings through eye-tracking metrics," *International Journal of Human-Computer Interaction*, pp. 1–20, 2025.
- [57] F. Li, C.-H. Lee, C.-H. Chen, and L. P. Khoo, "Hybrid data-driven vigilance model in traffic control center using eye-tracking data and context data," *Advanced Engineering Informatics*, vol. 42, p. 100940, 2019.
- [58] M. R. Harwood, *The Fourier analysis of saccadic eye movements*. University of London, University College London (United Kingdom), 2003.
- [59] Y.-J. Yoo, "Fault detection of induction motor using fast fourier transform with feature selection via principal component analysis," *International Journal of Precision Engineering and Manufacturing*, vol. 20, no. 9, pp. 1543–1552, 2019.
- [60] H. Subramanian, B. L. Ibey, W. Xu, M. A. Wilson, M. N. Ericson, and G. L. Coté, "An autocorrelation-based time domain analysis technique for monitoring perfusion and oxygenation in transplanted organs," *IEEE transactions on biomedical engineering*, vol. 52, no. 7, pp. 1355–1358, 2005.
- [61] D. V. Silagyi II and D. Liu, "Prediction of severity of aviation landing accidents using support vector machine models," *Accident Analysis & Prevention*, vol. 187, p. 107043, 2023.
- [62] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
- [63] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
- [64] Y. Zhang, S. Yang, G. Cauwenberghs, and T.-P. Jung, "From word embedding to reading embedding using large language model, eeg and eye-tracking," in *2024 46th Annual International Conference of the IEEE*

*Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–4.

- [65] A. R. Suhari, R. Hartanto, and S. Wibirama, “Improving performance of eye movements classification using cnn-transformer model,” in *2023 International Conference on Advanced Mechatronics, Intelligent Manufacturing and Industrial Automation (ICAMIMIA)*. IEEE, 2023, pp. 751–756.
- [66] P. Zhou, L. Chen, X. Dai, B. Li, and T. Chai, “Intelligent prediction of train delay changes and propagation using rvflns with improved transfer learning and ensemble learning,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7432–7444, 2020.
- [67] X. Yu, C.-H. Chen, and H. Yang, “Air traffic controllers’ mental fatigue recognition: A multi-sensor information fusion-based deep learning approach,” *Advanced Engineering Informatics*, vol. 57, p. 102123, 2023.
- [68] L. Mengtao, L. Fan, X. Gangyan, and H. Su, “Leveraging eye-tracking technologies to promote aviation safety—a review of key aspects, challenges, and future perspectives,” *Safety science*, vol. 168, p. 106295, 2023.
- [69] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.
- [70] S. He, R. Bao, J. Li, P. E. Grant, and Y. Ou, “Accuracy of segment-anything model (sam) in medical image segmentation tasks,” *CoRR*, 2023.
- [71] D. M. O’Shea and R. A. Fieo, “Individual differences in fluid intelligence predicts inattentive blindness in a sample of older adults: A preliminary study,” *Psychological research*, vol. 79, pp. 570–578, 2015.
- [72] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [73] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.
- [74] Q. Zhu, B. Hou, T. S. Mathai, P. Mukherjee, Q. Jin, X. Chen, Z. Wang, R. Cheng, R. M. Summers, and Z. Lu, “How well do multi-modal llms interpret ct scans? an auto-evaluation framework for analyses,” *arXiv preprint arXiv:2403.05680*, 2024.
- [75] J. Gripenkoven and S. Dietsch, “Gaze direction and driving behavior of drivers at level crossings,” *Journal of Transportation Safety & Security*, vol. 8, no. sup1, pp. 4–18, 2016.
- [76] H. Zhang and M. Xu, “Modeling temporal information using discrete fourier transform for recognizing emotions in user-generated videos,” in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 629–633.
- [77] K. Harezlak and P. Kasprowski, “Searching for chaos evidence in eye movement signals,” *Entropy*, vol. 20, no. 1, p. 32, 2018.
- [78] D. D. Clarke, P. Ward, C. Bartle, and W. Truman, “The role of motorcyclist and other driver behaviour in two types of serious accident in the uk,” *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 974–981, 2007.
- [79] D. Guo, Z. Zhang, B. Yang, J. Zhang, H. Yang, and Y. Lin, “Integrating spoken instructions into flight trajectory prediction to optimize automation in air traffic control,” *Nature Communications*, vol. 15, no. 1, p. 9662, 2024.



**Zhimin Li** received the B.Sc. degree in logistics management from Guangzhou University, China, and the M.Sc. degree in operations management from Shenzhen University, China. She is currently pursuing a Ph.D. degree with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong.

Her research interests include ergonomics in air traffic control, human-computer interaction, and data-driven safety monitoring.



**Fan Li** received the B.Eng. Degree in School of Aeronautic Science and Engineering, Beihang University, and M.Sc. Degree and Ph.D. Degree in School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore in 2015 and 2020. She is currently an Assistant Professor in the Department of Aeronautical and Aviation Engineering (AAE), at The Hong Kong Polytechnic University, HKSAR. Before joining AAE, she was a Research Fellow in Design and Human Factors Laboratory, the School of Mechanical and Aerospace

Engineering and Fraunhofer IDM@NTU center, Nanyang Technological University, Singapore.

Her research interests include safety management systems, human-centered design, intelligent transportation systems, and sustainable human-computer interaction.



**Gangyan Xu** (Member, IEEE) received the B.S. degree in automation and the M.E. degree in systems engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively, and the Ph.D. degree in systems engineering from The University of Hong Kong, Hong Kong, in 2016.

He is currently an Assistant Professor with The Hong Kong Polytechnic University, Hong Kong. Prior to that, he was an Assistant Professor with the Harbin Institute of Technology, Shenzhen, China; a Research Fellow with Nanyang Technological University, Singapore; and a Research Assistant with the City University of Hong Kong, Hong Kong. His research interests include data-driven optimization and control, intelligent transportation systems, resilient engineering, and emergency management.

Dr. Xu is an Editorial Board Member of *Advanced Engineering Informatics* and a Special Corresponding Expert of *Frontiers of Engineering Management*.



**Donglin Li** received the B.S. degree in Transportation from Civil Aviation Flight University of China University, Guanghan, China, in 2020, and the M.Sc. degree in Aviation Engineering from The Hong Kong Polytechnic University, Hong Kong, in 2024. He is currently pursuing a Ph.D. degree with the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong. From 2020 to 2022, he was an Approach Air Traffic Controller in Ningxia Branch of Northwest Air Traffic Management Bureau of Civil Aviation of

China.

His research interests include air traffic control, human-computer interaction, and human-centred design.