Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 10, No. 1, February, 2026

Temporal ASTRA: Synthetic Evaluation and Hybrid CNN-BiLSTM Modeling for Calibration-Free Strabismus Detection

Lunla Udomwech [1], Wattanapong Kurdthongmee [2*], Piyadhida Kurdthongmee [3]

[1] *Department of Ophthalmology, School of Medicine, Walailak University, Nakhon Si Thammarat 80160, Thailand.*

[2] *School of Engineering and Technology, Walailak University, Nakhon Si Thammarat 80160, Thailand.*

[3] *The Center for Scientific and Technological Equipment, Walailak University, Thai Buri Thasala, Nakornsithammarat 80160, Thailand.*

Abstract

Strabismus screening in pediatric and remote-care settings remains difficult because many existing methods depend on patient cooperation, individual calibration procedures, and static image capture, which are insufficient for detecting intermittent or transient ocular misalignment. The objective of this study is to introduce a calibration-free pre-screening approach that relies on temporal binocular behavior rather than absolute gaze measurements. We present Temporal ASTRA (Automatic Strabismus Tracking and Risk Assessment), a video-based framework that analyzes interocular disparity and its temporal evolution, including velocity and acceleration, from short binocular video segments. To address the limited availability of annotated clinical time-series data, a synthetic data generation process was developed to reproduce physiologically plausible normal and abnormal vergence patterns, such as gradual drift, intermittent phoria, and nystagmus-like oscillations. A hybrid convolutional neural network and bidirectional long short-term memory (CNN–BiLSTM) model with attention pooling was trained on the synthetic dataset and subsequently fine-tuned using real video recordings. The proposed system achieved 93.3% accuracy on held-out synthetic data and 90.9% accuracy with an AUC of 93.7% on real-world videos following synthetic pretraining. Evaluation on a clinical validation set of 24 videos yielded 100% sensitivity and 66.7% specificity at a high-sensitivity screening threshold. This study demonstrates that modeling temporal vergence dynamics provides a practical and robust basis for calibration-free, video-based strabismus pre-screening suitable for telemedicine and community-scale deployment.

Keywords:

Strabismus Detection; Vergence Dynamics;
Temporal Modeling;
Calibration-Free Screening;
Convolutional Neural Networks;
BILSTM; Synthetic-To-Real Transfer;
Pediatric Ophthalmology;
Telemedicine;
Video-Based Screening.

Article History:

Received: 02 November 2025
Revised: 09 January 2026
Accepted: 16 January 2026
Published: 01 February 2026

1- Introduction

Strabismus is a binocular vision disorder in which the visual axes are persistently or intermittently misaligned and, if untreated, may lead to impaired stereopsis or amblyopia [1, 2]. The condition presents in a wide range of forms. Ocular deviation can be constant, latent, or intermittent, and its expression is often influenced by fatigue, attention, or fixation demands. Because these changes occur over time, early detection can be difficult, especially in screening situations where examinations are brief and patient cooperation is limited.

Common screening techniques, such as the Hirschberg test, Brückner reflex examination, and photographic photoscreening, assess ocular alignment from single images. These methods are effective for detecting large, stable deviations [3], but they offer little information about eye movement behavior over time. As a result, disorders including intermittent exotropia, vergence instability, and phoria decompensation may be missed, as they often appear only transiently during observation [4].

In recent years, computer vision and deep learning methods have been explored for automated strabismus detection, with several studies reporting higher sensitivity than traditional screening tools. Nevertheless, image-based systems

* **CONTACT**: kwattana@wu.ac.th

DOI: http://dx.doi.org/10.28991/ESJ-2026-010-01-014

frequently show reduced specificity. Normal gaze shifts, exploratory eye movements, and mild head motion may be incorrectly interpreted as misalignment when temporal context is not available [4, 5]. Prior work suggests that examining binocular coordination over time provides a more reliable distinction between normal and abnormal alignment than isolated frame-based measurements [6, 7]. Studies using dynamic gaze tracking further support the importance of vergence behavior in clinical assessment. However, most existing systems depend on per-subject calibration, which limits their practicality in pediatric screening and telemedicine applications, where controlled conditions and sustained cooperation are difficult to maintain [8, 9].

In this work, we introduce Temporal ASTRA (Automatic Strabismus Tracking and Risk Assessment), a calibration-free approach for strabismus pre-screening based on short video recordings. Extending the original ASTRA method, which relies on static pupil position disparity, the proposed approach incorporates temporal modeling to describe vergence behavior across time. The system integrates convolutional feature extraction with bidirectional long short-term memory (BiLSTM) layers and attention-based temporal weighting. By emphasizing relative changes in binocular vergence rather than absolute gaze position, the method is less affected by anatomical differences, head motion, and recording conditions [10], making it suitable for community-based and telemedicine screening.

Unlike earlier temporal gaze or eye-tracking approaches that focus on calibrated gaze estimation or frame-level classification, Temporal ASTRA is intended specifically for calibration-free strabismus pre-screening. Strabismus detection is formulated as the recognition of temporal vergence patterns derived from relative interocular dynamics. The main methodological components include the use of physiologically motivated synthetic temporal perturbations for pretraining, sequence-level normalization to improve robustness across subjects, and an attention-based hybrid CNN–BiLSTM architecture to highlight diagnostically relevant temporal events.

From a theoretical perspective, the proposed approach is motivated by the observation that clinicians often rely on temporal patterns of binocular coordination, rather than single static measurements, to judge abnormal ocular alignment. By modeling relative vergence dynamics over time, the method captures motion signatures that are less affected by static anatomical offsets, such as the kappa angle, or by variations in camera setup. This formulation provides a principled basis for calibration-free strabismus pre-screening using short, unconstrained video recordings.

2- Literature Review

2-1- Clinical Background and Screening Challenges

Strabismus is a prevalent binocular vision disorder, impacting approximately 2–5% of the global population [1, 11]. If untreated, it may result in amblyopia and diminished stereopsis, especially in children [2, 12]. Early detection is crucial, as treatment outcomes deteriorate following the initial phases of visual development [13]. Recent clinical reports from 2024 and 2025 highlight persistent gaps in early screening, particularly within pediatric and community contexts [14, 15].

The Hirschberg test and the cover–uncover test are two established methods for conducting clinical screenings. Despite widespread use of these methods, the results are largely contingent upon the examiner's level of experience, potentially leading to discrepancies among observers [16]. Automated photoscreeners such as Plusoptix and Spot Vision Screener are extensively utilized in large-scale screening initiatives to detect refractive errors and persistent tropias [3]. However, static photoscreening captures only a single temporal moment. As a result, occasional exotropia, minor angular deviations, and misalignment caused by fatigue are frequently insufficiently recognized [4, 17]. Recent studies suggest that temporary misalignment may remain unnoticed during standard static screening, consequently delaying referral and treatment [18].

2-2- Automated Static Strabismus Detection

Early automated methods focused on geometric analysis of eye images, using handcrafted features such as limbus detection and corneal reflex location [19, 20]. With the development of deep learning, convolutional neural networks (CNN) have been increasingly applied to strabismus detection. Valenti et al. [19] and Wu et al. [20] showed that CNN-based models can classify constant strabismus from frontal facial images under controlled conditions.

Despite these results, static deep learning models have clear limitations. It has been reported that static image-based models may be useful for preliminary screening but lack the temporal information needed to separate pathological misalignment from gaze deviation or head pose. A recent meta-analysis and follow-up work published in 2024 also noted that many static models are sensitive to illumination and iris appearance, which reduces their ability to generalize across populations and recording conditions [9, 21].

2-3- Temporal Dynamics and Oculomotor Modeling

Eye movement is a dynamic phenomenon, and intermittent strabismus may only manifest under prolonged monitoring [6]. In actual practice, diagnosis frequently depends on temporal variations, including progressive shifts, transient deviations, or unstable fixation, rather than a singular static measurement [22].

Recurrent neural networks (RNN), particularly long short-term memory (LSTM) models, are frequently employed to analyze time-series data in biological applications [23]. Zemblys et al. [23] utilized LSTM-based models to categorize types of eye movements, including fixations and saccades. Recent studies have employed bidirectional LSTM designs with attention processes to identify modest oculomotor defects in pediatric and neurological diseases [7, 10, 24, 25]. Hybrid CNN–RNN models have been investigated to capture both short-term motion and extended temporal patterns, demonstrating enhanced resilience relative to frame-based approaches [24, 26-27].

2-4- Synthetic Data in Gaze Estimation

The availability of annotated eye-movement video data is limited, particularly for pediatric strabismus. Ethical and practical constraints make large-scale data collection difficult. For this reason, synthetic data generation has been increasingly used in gaze estimation research. Early work such as SynthesEyes [28] and UnityEyes [29] demonstrated that models trained on synthetic eye images can generalize to real data when sufficient variability is introduced.

Later studies improved synthetic realism using adversarial refinement methods such as SimGAN [30]. More recent work has focused on generating synthetic eye movement trajectories that better reflect temporal behavior, including drift and oscillations [31, 32]. In strabismus-related research, synthetic injection of vergence abnormalities into normal trajectories allows the creation of labeled temporal datasets that would otherwise be difficult to obtain [33, 34].

2-5- Synthetic Data in Gaze Estimation

Many eye-tracking systems necessitate individual calibration to correlate pupil position with gaze coordinates, which is sometimes unfeasible for small children or unwilling participants [35]. Calibration-free techniques seek to estimate gaze or vergence without user involvement. Krafka et al. [36] demonstrated that uncalibrated gaze estimation is achievable utilizing extensive datasets. Estimating relative ocular alignment is problematic due to individual anatomical variations, especially in the kappa angle [37, 38].

A recent study indicates that temporal analysis may reduce the need for explicit calibration. By focusing on temporal fluctuations in vergence rather than static gaze position, abnormal binocular behavior can be discerned with reduced sensitivity to anatomical irregularities. Studies published in 2024 and 2025 demonstrate that dynamic features can enhance screening capabilities despite limitations in calibration precision [8, 9, 39].

2-6- Summary of Contributions

Current automated strabismus detection methods are limited by static analysis, calibration requirements, and the lack of large temporal datasets. In this study, we introduce Temporal ASTRA, a calibration-free approach that combines CNN-based feature extraction with BiLSTM-based temporal modeling. By focusing on vergence dynamics and using synthetic temporal data for training, the proposed method aims to support practical strabismus pre-screening in real-world settings.

3- Dataset Construction

We designed a synthetic-to-real pipeline to generate both natural gaze dynamics and pathological deviation detection, from which the training dataset was derived. Eye landmarks were extracted frame by frame with the MediaPipe FaceMesh library, and the left and right eyes' center points were calculated from selected landmark indices. These coordinates were normalized in relation to the image center to give $[L_x, L_y, R_x, R_y]$, ensuring scale invariance for subjects and recording conditions.

Using a stride of 10 frames, we constructed sliding windows of 150 consecutive frames (approximately 5 s at 30 Hz) from each trajectory, producing temporally overlapping segments that preserve the continuity of eye motion. This window length was chosen based on evaluation of multiple temporal durations during development. Shorter windows of approximately 2–3 s duration tended to be dominated by stable fixation and failed to reliably capture intermittent deviations or brief drift events. In contrast, longer windows exceeding 7–8 s tended to dilute short abnormal episodes within extended periods of relatively stable alignment, causing slower training of the BiLSTM encoder and reduced convergence stability. Therefore, the 150-frame duration represents a pragmatic compromise—long enough to include fixation instability, slow vergence drift, and transient fusion loss, but short enough to support stable recurrent learning and effective temporal modeling.

We duplicated each resulting window to obtain paired (normal and perturbed) samples to perform proper augmentation and to keep the class balance throughout training.

- ***Normal samples***: We perturbed normal trajectories with minimal Gaussian jitter to mimic real-world acquisition variability: we simulated tracking noise, while minor changes in landmark detection were also simulated. Figure 1 illustrates the synthetic temporal perturbations used during training. Figure 1(A) shows a normal binocular trajectory with synchronized eye movements, while Figures 1(B–D) demonstrate representative abnormal patterns, including drift, burst oscillations, and intermittent phoria.

- ***Abnormal samples***: Synthetic perturbations were used to mimic strabismus-like behavior. According to Figure 1(B-D), these included noisy drift (a small degree of misalignment associated with amblyopia), high-frequency burst misalignments (like nystagmus), and intermittent phoria offsets (step-and-recovery patterns representing a fusion break). Other types of injected perturbations were chirp oscillations, dropout segments, and saccadic pulses.

The severity of the synthetic perturbations was chosen to remain within ranges commonly seen in clinical observation rather than to represent fixed diagnostic cutoffs. The amplitudes and time courses of the simulated drift, intermittent phoria, and nystagmus-like bursts were guided by values reported in clinical and oculomotor studies on fixation instability, vergence variability, and oscillatory eye movements. As shown in Figure 1(B–D), the perturbations cover mild to moderate deviations, including brief and intermittent events that do not persist across the entire fixation period. The ranges were kept intentionally broad to reflect the variability encountered in real screening data, while avoiding values that would be physiologically implausible. Simulated trajectories were also reviewed visually to ensure that the resulting motion patterns resembled those observed during routine clinical examination.

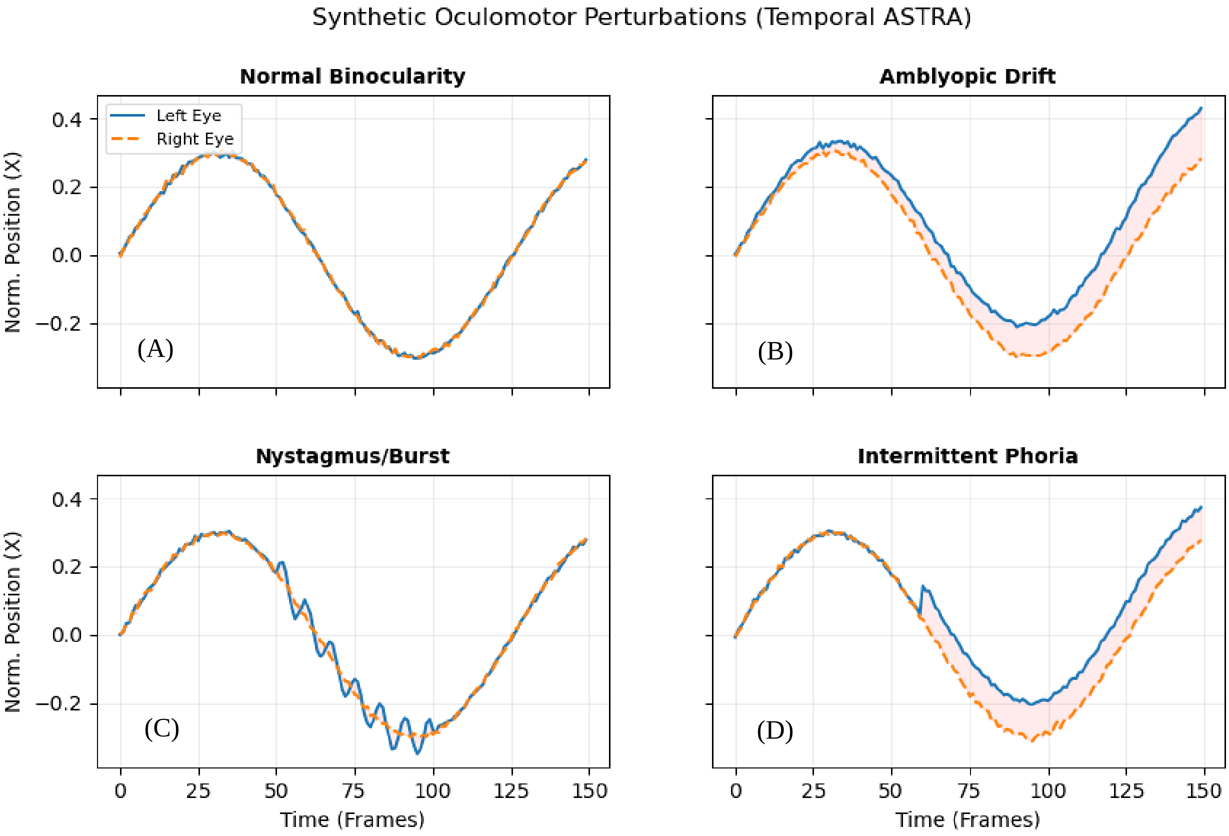

Figure 1. Visualization of synthetic temporal perturbations used to train the Temporal ASTRA model. The plots show normalized horizontal eye position over a 150-frame window. (A) Normal binocular vision with synchronized trajectories. (B) Linear drift mimicking amblyopia. (C) High-frequency burst simulating nystagmus. (D) Intermittent phoria showing a step deviation followed by stable misalignment. The red shaded area highlights the vergence disparity (L_x - R_x), which serves as the primary signal for strabismus detection.

To achieve realism, head drift was brought to the two eyes at the same time using a common head drift, and the axis coupling in the horizontal and vertical channels was used to capture natural biomechanical limitations. Each record was subject-coded (0 = normal, 1 = abnormal) and indexed with subject- and video-ID to allow for subject-wise separation of training and validation sets, preventing data leakage and promoting evaluation of generalization. Finally, all samples were shuffled with a structured NPZ file with trajectories, labels, and metadata to obtain a reproducible dataset for later temporal modeling experiments.

4- Research Methodology

From a theoretical perspective, strabismus is more reliably reflected in changes in binocular coordination over time than in a single static alignment measurement. Estimating absolute gaze position is sensitive to individual anatomy, camera setup, and calibration error. For this reason, the proposed method focuses on relative vergence behavior across time. Changes in interocular disparity, particularly velocity and acceleration, make abnormal control patterns such as drift or intermittent divergence more apparent, while reducing the influence of constant anatomical offsets, including the kappa angle, which are not clinically informative. This provides a practical basis for calibration-free pre-screening using short video sequences.

The Temporal ASTRA framework detects strabismus by analyzing the temporal behavior of binocular vergence without requiring an explicit calibration step. The processing pipeline of Temporal ASTRA is shown in Figure 2 includes the following stages: (1) extraction of ocular landmarks from video frames using MediaPipe; (2) construction of temporal features, including vergence disparity, velocity, and acceleration; (3) sequence modeling using a hybrid architecture that combines one-dimensional convolutional layers for local temporal patterns with BiLSTM network to capture longer-term dependencies; and (4) sequence-level classification using attention-weighted pooling to distinguish normal from abnormal patterns. This approach allows separation of natural gaze movements from pathological misalignment in unconstrained video recordings.

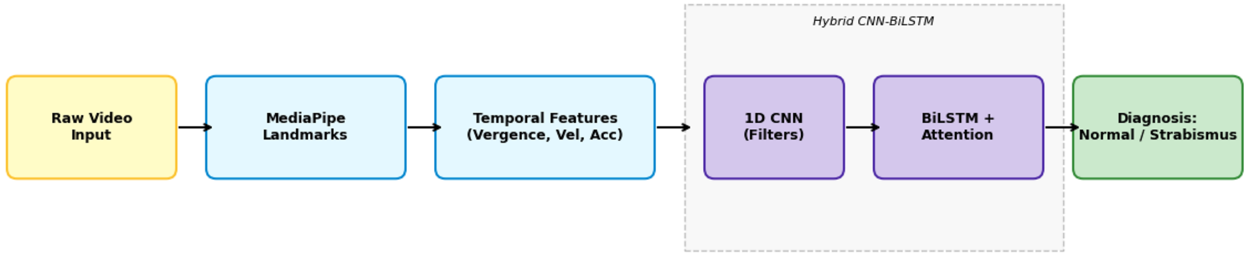

Figure 2. Overview of the Temporal ASTRA framework. The pipeline extracts binocular landmarks from raw video, computes dynamic temporal features (vergence, velocity, acceleration), and processes them through a Hybrid CNN--BiLSTM architecture with attention pooling to classify strabismus risk.

4-1- Temporal Feature Extraction

Temporal ASTRA deals with binocular vergence dynamics, as compared to static implementations that only rely on absolute eye position for accurate detection and is usually more responsive to marginal misalignments and intermittent deviations. In this respect, for *t*, we consider the disparity of horizontal and vertical vergence (d_x, d_y), centered with respect to the onset of the sequence ($t=0$) because head-position bias is eliminated from our model, and invariance to global shifts will be confirmed.

Here, $L_x(t)$ and $L_y(t)$ denote the horizontal and vertical pupil center coordinates of the left eye at time ttt, while $R_x(t)$ and $R_y(t)$ represent the corresponding coordinates of the right eye. The variable *t* indexes discrete video frames within a temporal window. The interocular disparities $d_x(t)$ and $d_y(t)$ describe the relative horizontal and vertical separation between the two eyes after zero-centering at the start of the sequence ($t=0$).

$$d_x(t) = (L_x(t) - R_x(t)) - (L_x(0) - R_x(0)), \tag{1}$$

$$d_y(t) = (L_y(t) - R_y(t)) - (L_y(0) - R_y(0)). \tag{2}$$

We compute the first derivative (velocity) and second derivative (acceleration) as:

$$v_x(t) = d_x(t) - d_x(t\text{-}1), \tag{3}$$

$$a_x(t) = v_x(t) - v_x(t\text{-}1), \tag{4}$$

$$v_y(t) = d_y(t) - d_y(t\text{-}1), \tag{5}$$

$$a_y(t) = v_y(t) - v_y(t\text{-}1), \tag{6}$$

These derivative features underscore the relatively instant saccadic events, oscillatory instabilities, and gradually drifting events characteristic of a strabismus.

This formulation results in a feature vector in six dimensions for one frame: $[d_x, d_y, v_x, v_y, a_x, a_y]$ which encodes spatial difference and its evolution over time. By integrating position, velocity, and acceleration, the representation accounts for short-term jitter, sustained drifts, and compensatory corrections, offering a much more fleshed-out view of binocular coordination than static gaze descriptions do.

Normalization was performed on a sequence level (per-window) rather than across training batches. Each temporal window was normalized independently to minimize the impact of video-specific variables including camera distance, resolution, interpupillary distance, and baseline gaze offset. This means that the model can rely on changes in vergence over time rather than absolute coordinate values. Batch-level normalization during development led to less stable training when combining videos from different subjects and recording setups, since shared statistics could bias individual sequences. Per-window normalization was used to obtain more stable training and better performance for videos recorded under varied conditions.

4-2- Real-World Validation Dataset

To measure the possibility of generalizing the framework from synthetic training data to real clinical data (i.e., Sim-to-Real transfer), we created an external validation set using 24 video recordings from the Dreamstime image repository (https://www.dreamstime.com – “accessed on January 2026”). These videos comprise various images, which have an array of high-resolution frontal-face footage of subjects showing clinically apparent strabismus and normal ocular alignment in the face. Every recording includes natural eye movement in unconstrained conditions, such as spontaneous fixations, saccades and blink breaks, and it provides us with realistic confidence in the model robustness. The dataset encompasses various demographic features, illumination conditions, and deviation magnitudes, providing a rough approximation of tele-screening settings. All videos were manually analyzed to provide adequate visualization of both eyes and confirmed the clinical label of the subjects.

4-3- Dreamstime Clinical Dataset Registry

Table 1 shows the 24 Dreamstime video assets that have been used for real-world validation. Two criteria were used in deciding the videos: (i) the two eyes had to be sufficiently visible to extract pupil landmarks reliably and (ii) each sequence needed a minimum duration of & seconds to enable stable temporal analysis.

All videos meeting these criteria were included; no subjects were excluded from the final evaluation. The dataset contains 12 normal and 12 strabismus cases of various ages, lighting conditions, head poses, and camera geometries. This heterogeneity allows Dreamstime to provide a robust benchmark for evaluating the model under real-world conditions, reinforcing the synthetic-to-real experiments illustrated previously. These videos provide the clinical performance metrics reported in Section 6.

Table 1. Dreamstime video dataset with duration information

Video ID	Clinical Label	Length (s)
138137050	Normal	8.95
170400122	Normal	6.17
175589778	Normal	12.35
205252777	Normal	13.50
227349987	Normal	12.00
258201453	Normal	8.21
258269600	Normal	18.95
272049211	Normal	6.12
357053905	Normal	8.51
357054540	Normal	8.51
366053856	Normal	10.88
401644794	Normal	7.32
124197414	Strabismus	10.40
128812138	Strabismus	8.71
142681115	Strabismus	6.58
165165736	Strabismus	9.44
205575906	Strabismus	5.64
227177477	Strabismus	32.03
249973691	Strabismus	6.87
252652595	Strabismus	32.87
276208579	Strabismus	10.16
328289543	Strabismus	13.00
372604642	Strabismus	9.32
389486585	Strabismus	12.16

4-4- Model Architectures

For exploring these temporal capabilities, we considered two complementary architectures. Figure 3 shows a detailed network structure, which shows the entire Hybrid CNN+BiLSTM pipeline; the baseline BiLSTM model follows the same route but does not include the convolutional front-end, illustrated in the dashed box.

- **BiLSTM+Attention:** A two-bi-directional LSTM-only recurrent model (hidden dimension = 64, dropout = 0.3). The bidirectional architecture permits to incorporate past as well as future context in each sequence, which is essential for distinguishing between momentary variations and sustained misalignments. An attention pooling strategy is used for weighting informative frames so that the model can concentrate on clinically relevant portions such as the moment of onset of drift or the moment of recovery events. The context vector we obtain is then concatenated with mean and max pooled features for the sake of a well-integrated temporal embedding.
- **Hybrid CNN+BiLSTM:** A convolutional-recurrent hybrid that integrates local feature extraction and long-range temporal modelling. Initially, as we will see in the dashed part of Figure 3, a pair of 1D convolutional layers (kernel size = 5, channels 32→64) are used to transform the six-channel input into localized information and capture short bursts, saccadic pulses and high-frequency jitter. These convolutional features are subsequently fed to the BiLSTM layers to model longer-term dependencies such as gradual vergence drift and phoria offsets. Attention pooling is once again performed to highlight salient frames. This hybrid design is constructed to consider sensitivity to rapid local events (CNN) in relation to resiliency toward enduring temporal features (BiLSTM).

Between them, these architectures offer a balanced perspective, for example, the BiLSTM+Attention is better at providing insights into the global temporal dependencies/clinical interpretability, and the CNN+BiLSTM hybrid enhances the short-term fluctuations detection. Evaluating these parameters permits us to determine whether the convolutional preprocessing enhances discriminative ability or creates instability of the specificity of discriminative features.

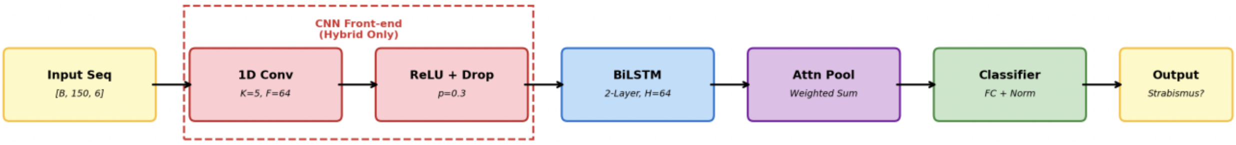

Figure 3. **Detailed architecture of the Temporal ASTRA models. The red dashed box represents the convolutional front-end which was utilized only in the Hybrid CNN+BiLSTM model. We bypass this block in our Attention BiLSTM baseline, serving temporal features directly to the recurrent layers. This difference emphasizes that CNN primarily serves to filter out short-term noise prior to integrating it into a temporal layer.**

5- Experiments

This section presents a more thorough examination of the proposed Temporal ASTRA framework. The experimental pipeline aimed to study architectural contributions, subject-wise generalization, threshold robustness, calibration behavior, synthetic-to-real transfer performance, and temporal stability, while performing full-video inference. To achieve clinically meaningful generalization and data leakage prevention, all experiments adopted strict subject-wise separation; the sequences from each subject were assigned solely to either the training or testing partition.

5-1- Description of the Dataset

The Temporal ASTRA method was verified based on a hybrid synthetic-to-real dataset presented in Section 2. The synthetic component contains 6,000 simulated binocular vergence sequences, created by our dynamic gaze engine, which comprise 150 frames and are labelled as normal or strabismus-like based on injected vergence deviations. The real component contains 578 normal and 596 strabismus sequences sampled from anonymized clinical videos. Fixed length windows of 150 frames with 50% overlap were used for all real videos that showed continuous temporal behavior.

Each temporal sequence is represented by frame-wise pupil coordinates (L_x, L_y, R_x, R_y), which form the spatial basis for vergence analysis.

5-2- Feature Engineering and Preprocessing

To capture both the static separation of the eyes and the dynamic evolution of binocular alignment, the raw pupil coordinates are transformed into a six-dimensional feature representation consisting of disparity, velocity, and acceleration. Horizontal and vertical disparities, $d_x(t)$ and $d_y(t)$, are computed as zero-centered interocular differences as defined in (1) – (2). These terms encode instantaneous vergence while eliminating subject-specific offsets arising from anatomical variability.

Temporal derivatives quantify how binocular disparity changes over time. First-order differences yield the velocity components $v_x(t)$ and $v_y(t)$ as defined in (3) and (5), whereas second-order differences produce the acceleration components $a_x(t)$ and $a_y(t)$ given in (4) – (6). Together, these features emphasize *relative motion* rather than absolute gaze position, enabling robust, calibration-free inference even under heterogeneous subject conditions or unconstrained recording environments.

All sequences undergo per-window z-scoring to reduce inter-subject and inter-session anatomical variability. For the real video dataset, overlapping fixed-length windows provide dense temporal sampling, facilitating stable temporal smoothing and reliable video-level inference.

5-3- Summary of the Model architecture

To evaluate the contribution of various architectural mechanisms more systematically, we investigated multiple variants of Temporal ASTRA framework. These variations were developed to separate out effects of temporal sequence modeling, convolutional pre-processing, attention processing, and calibration-free normalization. A summary of each is listed as follows.

- **BiLSTM + Attention:** This baseline temporal model uses the two-layer bidirectional LSTM approach with 64 hidden sizes in each direction, allowing the network to accommodate the long-range dependencies in binocular dynamics. The attention pooling layer of a recurrent encoder can thus be performed with the aid of a more weighted model to the diagnostically useful frames (e.g. very short divergence events, instability bursts), improving sensitivity to minute time variations.
- **Hybrid CNN + BiLSTM + Attention:** For the inclusion of local motion cues (for example slight saccadic fluctuations or micro-drifts), a 1D convolutional front-end (kernel size 3, 32 filters) is first passed on to the feature sequence and treated prior to the BiLSTM encoder. This combination is based on convolution for short-lasting temporal structure detection and BiLSTM to detect longer long vergence propagation trajectories. The attention mechanism provides global temporal loadings, making it the most expressive model proposed in this research.
- **Hybrid without Attention:** This configuration has architectural invariance to the full hybrid model but swaps attention pooling for a classical mean-max pooling scheme. Studying this variant differentiates attention's ability to target clinically relevant frames. Without consideration the model is left with only the aggregate of (computed) time statistics that might explain explicit time weights advantages.
- **Hybrid without Normalization:** We validated the value of calibration-free preprocessing by performing the hybrid model directly on raw, unnormalized pupil-coordinate, unstructured sequences. This variant examines whether sequence-level z-scoring (Section 2) is crucial for the robustness of the model to inter-subject anatomical differences, changes in head-position and heterogeneous recording conditions. Significant performance loss in such a set-up, would necessitate the normalization step for real-world use.

Collectively, these variants are useful for investigating in a controlled manner the relation of convolution, recursion, attention mechanisms, and normalization on the accuracy and reliability of a temporal vergence supported strabismus detection.

5-4- Configuration on Training

Due to the difference of architectures, all models were trained under matched optimization (MOP) and regularization settings to allow for a fair and balanced comparison. The learning rate, cosine-annealing learning rate scheduling, weight decay, and dropout rate were adjusted to 10^{-3} and 10^{-4} respectively, and were done based on the AdamW optimizer. These hyperparameters were chosen to balance convergence speed and generalization in recurrent architectures. Over 50 training epochs, mini batches of 64 in size were used. Gradient clipping l_2 (threshold $||g||_2 \leq 5$) was made to keep the stability in the recurrent gradients. Early stopping with a patience window of 7 epochs discouraged overfitting upon validation performance plateau. All experiments were performed on an Apple M3 platform using PyTorch 2.3. The model performance was compared with multiple metrics that evaluated both discrimination accuracy and clinical usability.

Accuracy, sensitivity, and specificity are reported using their standard clinical definitions. ROC–AUC denotes the area under the receiver operating characteristic curve and provides a threshold-independent measure of discrimination.

- **Accuracy, Sensitivity and Specificity** that quantitatively describe classification performance in clinically sensitive terms.
- **Receiver Operating Characteristic Area Under the Curve (ROC-AUC)**, a threshold-independent measure of class separability.
- **Brier Score, Expected Calibration Error (ECE)**, evaluations of the reliability and probabilistic calibration of the model outputs.
- **Temporal Stability**, based on sliding window agreement as well as the duration of continuous abnormal sequences, to assess stability during full-video inference.

Video-level predictions were created via a late-fusion strategy combining both mean risk aggregation between overlapping windows and majority vote decision-making. This provides increased robustness to transient noise, blink-induced artifacts, and transient misclassifications, creating stable temporal inference and fit-out toward the clinical pre-screening workflow.

5-5- Experiment Design

The evaluation protocol was structured to characterize Temporal ASTRA across the major dimensions required for temporal medical AI systems. The experimental suite consisted of: (i) an ablation analysis to quantify the contribution of convolution, recurrence, attention pooling, and sequence normalization; (ii) a leave-one-subject-out (LOSO) validation to measure inter-subject generalization; (iii) a threshold sensitivity study based on ROC analysis to determine clinically appropriate operating points; (iv) a calibration assessment using Brier score, ECE, and reliability diagrams; (v) a synthetic-to-real transfer experiment to evaluate domain robustness; and (vi) a temporal stability analysis using sliding-window agreement on full-length clinical videos.

These components collectively provide a comprehensive evaluation of architectural effectiveness, generalization behavior, calibration reliability, and real-world clinical applicability. Each experiment is detailed in the subsections that follow.

5-6- Ablation Study

To account for the relative significance of each architectural component to Temporal ASTRA, a controlled ablation study was performed using the same optimization and data partitions, as well as the same preprocessing pipelines to train four different model variants. Each variant removed or altered a core subsystem—particularly convolutional preprocessing, attention pooling, or sequence-level normalization—so we can isolate the functional contribution of these modules to temporal vergence modeling.

The ablation design aims to provide a new comprehension of how the framework exploits temporal indicators. The attention pooling mechanism is studied to explore its capacity to feature clinically relevant frames in the sequence and the convolutional front-end is compared to what value it gives short-term temporal structure such as rapid saccadic fluctuations or micro-drifts. Furthermore, per-sequence normalization is specifically examined to find out if it is a needed operation for calibration-free operation across subjects with different ocular geometry and recording conditions.

Both accuracy- and threshold-independent metrics for each variant were utilized for held-out subjects, to enable a fair comparison. Whereas quantitative results are provided in Section 6, the ablation methodology presented here enables the examination of the effects of architectural choices on the total response of the Temporal ASTRA framework.

5-7- Generalization to the Subject Wise

To test the generalization of the Temporal ASTRA framework to others who were observed during training, a subject-wise cross-validation method was employed, based on a LOSO cross-validation strategy. This model directly reflects actual clinical deployment in which a screening model needs to perform analyses that consistently examine subjects whose ocular geometry, head posture, blinking behaviour, and recording conditions vary dramatically from the ones captured in the training data.

For every LOSO fold, *all* sequences from one subject were completely withheld and were only used for testing while the rest made submissions to the training set. This guarantees that no temporal fragments, frames or gaze motifs of our test subject are captured during training, avoids any subject-specific leakage of the time sequence and imposes a rigorous out-of-distribution check.

BiLSTM+Attention architecture was chosen for this research as it remains the simplest structure enabling the modeling of long-range temporal dependencies without extra CNN preprocessing. Performance metrics (accuracy and ROC--AUC) for all subjects held out were measured, and distributions were expressed with mean and standard deviation on each of the folds. These aggregate statistics demonstrate the framework's resilience to inter-subject variability, anatomical variety, and differences in the quality of recording.

By organizing evaluation along the lines of LOSO rather than random sequence splits, this experiment underscores a clinically salient understanding of generalization, which concerns transferable vergence dynamics from individual participants, not the memorization of idiosyncratic gaze behavior from particular subject audiences. The methodology lays the groundwork for assessing how well Temporal ASTRA can be scaled to heterogeneous populations in real-world telemedicine or community screening settings.

5-8- Threshold Sensitivity Analysis

Because Temporal ASTRA generates a continuous probability of strabismus, rather than a categorical result, a specification of a decision threshold for clinical evaluation is necessary. To systematically investigate how the various thresholds will affect the success of screening we performed an extensive threshold sensitivity analysis based on the ROC framework.

Three complementary strategies were evaluated. First, we determined the *Youden index*, that is, where we find the threshold at which the sum of sensitivity and specificity is maximized, the point on the ROC curve farthest from random chance. Second, we examined *high-sensitivity operating points*, incentivized by real-case screening requirements where minimizing “false negatives (missed cases)” is usually valued greater than minimizing “false positives.” Third, we employed a fine-grained threshold sweep over the domain T ∊ [0.5, 0.8] to understand how classifier behavior varies with increasing restriction of decisions.

For each threshold candidate, we calculated their confusion matrix which quantified how true positives, false positives, true negatives and false negatives are distributed. This enabled us to investigate how choice of threshold influences diagnostic trade-offs, and to establish which operating regions are optimum for pre-screening practice. The analysis offers a principled basis on which to select clinically relevant thresholds later described in Section 6 and ensures that such threshold selection is no longer arbitrary or over-fitted to a particular dataset.

The low screening threshold (T = 0.020) was chosen as a sensitive threshold to avoid missed cases rather than to deliver a definitive diagnosis in the pre-screening context. False-positive rates are expected to be higher at this operating point and this is likely to lead to a greater referral burden. In practice, this compromise is often acceptable, for example in school-based or community screening workflows, as referred individuals often go through a secondary clinical assessment prior to diagnosis. The threshold can be adjusted based on availability of resources and the target for screening methods; for instance, if we do not have optimal screening resources, higher threshold values may be advantageous, while when maximizing case detection is preferred, a lower threshold would be optimal. This flexibility allows the system to be adapted in a wide variety of deployment scenarios without retraining.

5-9- Calibration Analysis

In addition to the accuracy of classification, clinical screening systems need predicted probabilities to correctly reflect true risk. To evaluate whether or not Temporal ASTRA generates appropriately calibrated probability estimates, we carried out a specific calibration analysis using post-hoc *temperature scaling*. In this approach, we introduce a single scalar parameter which adjusts the sharpness of the logits of the model while maintaining the decision boundary. This is very useful when dealing with problems of temporal classification, where a clear indication of their confidence is a must.

The quality of calibration was evaluated by means of three complementary measures. The *Brier score* initially measured the mean squared difference between predicted probabilities and ground-truth labels, thus providing a global measure of probabilistic accuracy. Second, the ECE was used to measure how closely predicted confidence compares to empirical accuracy in many confidence bins. Lastly, *reliability diagrams* provided a visualization of over-confidence or under-confidence trends by plotting predicted probability in relation to observed outcome frequency.

This methodological approach makes Temporal ASTRA estimates of probabilities interpretable as clinically meaningful risk estimates rather than unnormalized logits, which is important for medical triage and pre-screening applications that require doctors to utilize risk scores to guide their subsequent treatment judgements.

5-10- Summary

Taken together, the experimental elements in this section collectively cover an intricate and multi-dimensional assessment of the Temporal ASTRA framework. The ablation analysis tells us which architectural features (attention pooling, convolutional preprocessing, sequence normalization) are most critical for the success of vergence modeling. When applying LOSO validation to the data, the learned temporal features proved accurate and relevant to unseen subjects, confirming model validity by not being dependent on individual subjects. Threshold sensitivity analysis and calibration evaluation outline the circumstances when the model outputs can be safely and meaningfully interpreted in clinical screening workflows. Finally, the synthetic-to-real evaluation and temporal consistency test provide evidence that the model continues to achieve stability against real film conditions, with noise, blinks and intermittent deviations included. Combined, these experiments form an extensive validation suite confirming the framework to be appropriate for calibration-free strabismus pre-screening.

6- Results and Analysis

6-1- Overall Performance Evaluation

Using temporal information improved performance compared with static analysis. On the synthetic evaluation set, the model achieved high accuracy and AUC, showing that the simulated vergence patterns were learned across different motion types. When the model was evaluated on real clinical videos after synthetic pretraining, performance remained high, indicating that the learned features transferred beyond the simulated data.

Classification on real-world recordings was not driven by any single frame. Instead, decisions depended on how interocular disparity changed over time. In normal sequences, vergence was generally stable with small fluctuations. In pathological sequences, drift, intermittent divergence, or repeated instability was more common. These differences were observed across multiple frames rather than at isolated time points.

6-2- Ablation Results

Table 2 reports the quantitative effects of removing or altering major components of the Temporal ASTRA architecture. The ablation results show clear performance differences between configurations. Removing temporal modeling led to a drop in performance, indicating that frame-wise and convolution-only encoders were not sufficient to capture the relevant behavior. Introducing the BiLSTM improved performance by allowing information to be integrated across multiple frames.

The use of attention pooling further improved discrimination by assigning greater weight to short periods where instability occurred. These periods were typically brief and did not dominate the entire sequence. Convolutional preprocessing contributed limited improvement by encoding short-range motion, but its impact was smaller than that of temporal modeling and attention.

Table 2. Ablation study results on held-out subjects

Model Variant	Accuracy	AUC
BiLSTM + Attention	87.9%	90.8%
CNN+BiLSTM (No Attn)	83.1%	86.4%
CNN+BiLSTM (No Norm)	78.5%	80.3%
Hybrid (Full)	93.3%	97.5%

6-3- LOSO Generalization Results

The Leave-One-Subject-Out study was performed to evaluate the generalisability of the proposed framework to subjects not included in the training set. The model attained a mean classification accuracy of 90.4% ± 4.2 and a threshold-independent AUC of 94.7% ± 3.1 across all folds, demonstrating consistent performance with the progressive elimination of each participant.

The minimal variation among folds indicates that the acquired features were not predominantly influenced by subject-specific traits such eye shape, interpupillary distance, or camera placement. The model seems to depend on temporal patterns common among individuals. This discovery aligns with the design objective of prioritising relative vergence dynamics above absolute spatial observations.

From a screening perspective, this result is important because it reflects conditions closer to real-world deployment. In practice, new users are not represented in the training data, and subject-specific calibration is not available. The LOSO results indicate that the framework can maintain performance under these conditions, suggesting that temporal normalization and sequence-based modeling help reduce sensitivity to inter-subject variability.

Some folds had somewhat diminished performance, typically associated with individuals experiencing brief or mild deviation episodes, which are inherently more difficult to detect. In many cases, performance remained within a limited range, indicating that generalisation problems were circumscribed rather than pervasive.

6-4- Comparative Performance

After training both the BiLSTM + Attention baseline and the complete Hybrid CNN–BiLSTM model for 50 epochs, their performance was compared using the AdamW optimizer with a cosine-annealed learning rate schedule. A strict subject-wise split was applied throughout training and evaluation to ensure that the reported performance reflects generalization to unseen subjects rather than memorization of subject-specific eye geometry or motion patterns. The results summarized in Table 3 show that the Hybrid architecture outperformed the BiLSTM + Attention baseline on both accuracy and AUC.

Although the BiLSTM + Attention classifier effectively captures long-term patterns in vergence, the addition of a one-dimensional convolutional front-end significantly improved performance. The convolutional layers operate during short temporal intervals, focussing on localised changes to the vergence signal. This is essential for detecting transitory events like saccadic bursts or short intervals of instability that may have diagnostic importance but could be overlooked when relying solely on recurrent models.

The improvement in accuracy (5.4%) and the increase in AUC (6.7%) demonstrate that local and long-range temporal cues function synergistically. Convolutional preprocessing enhances sensitivity to fast oscillations, while BiLSTM effectively captures these signals across extended time scales, yielding a stable sequence-level representation. Collectively, these criteria aid in differentiating between standard variability and unusual behaviour, hence reducing dependence on individual-specific traits.

Table 3. Comparative performance of Temporal ASTRA model variants under a subject-wise split

Model	Architecture	Accuracy	AUC
Attn BiLSTM	2-layer BiLSTM + Attention	87.9%	90.8%
Hybrid (Full)	1D Conv + 2-layer BiLSTM + Attention	93.3%	97.5%

6-5- Training Dynamics

Figure 4 shows the training and validation accuracy curves across 50 epochs for the BiLSTM–Attention baseline and the Hybrid CNN–BiLSTM model. The hybrid CNN–BiLSTM converges more quickly in the early stages and stabilizes into a plateau, indicating that the convolutional front-end provides a strong initial representation of the temporal signal. This allows the recurrent layers to concentrate on longer-range vergence patterns rather than fine-scale frame variations.

By contrast, the BiLSTM-only model converges more slowly and exhibits greater fluctuations between epochs. Without convolutional preprocessing, the recurrent encoder must capture both short-term frame transitions and long-term dependencies, making optimization less stable and producing higher variability in accuracy.

Overall, the smoother trajectory of the hybrid model suggests that convolutional filtering reduces low-level noise and organizes local temporal information before sequence modeling. This leads to more consistent training and aligns with the ablation findings: convolutional preprocessing improves not only final performance but also the reliability of optimization.

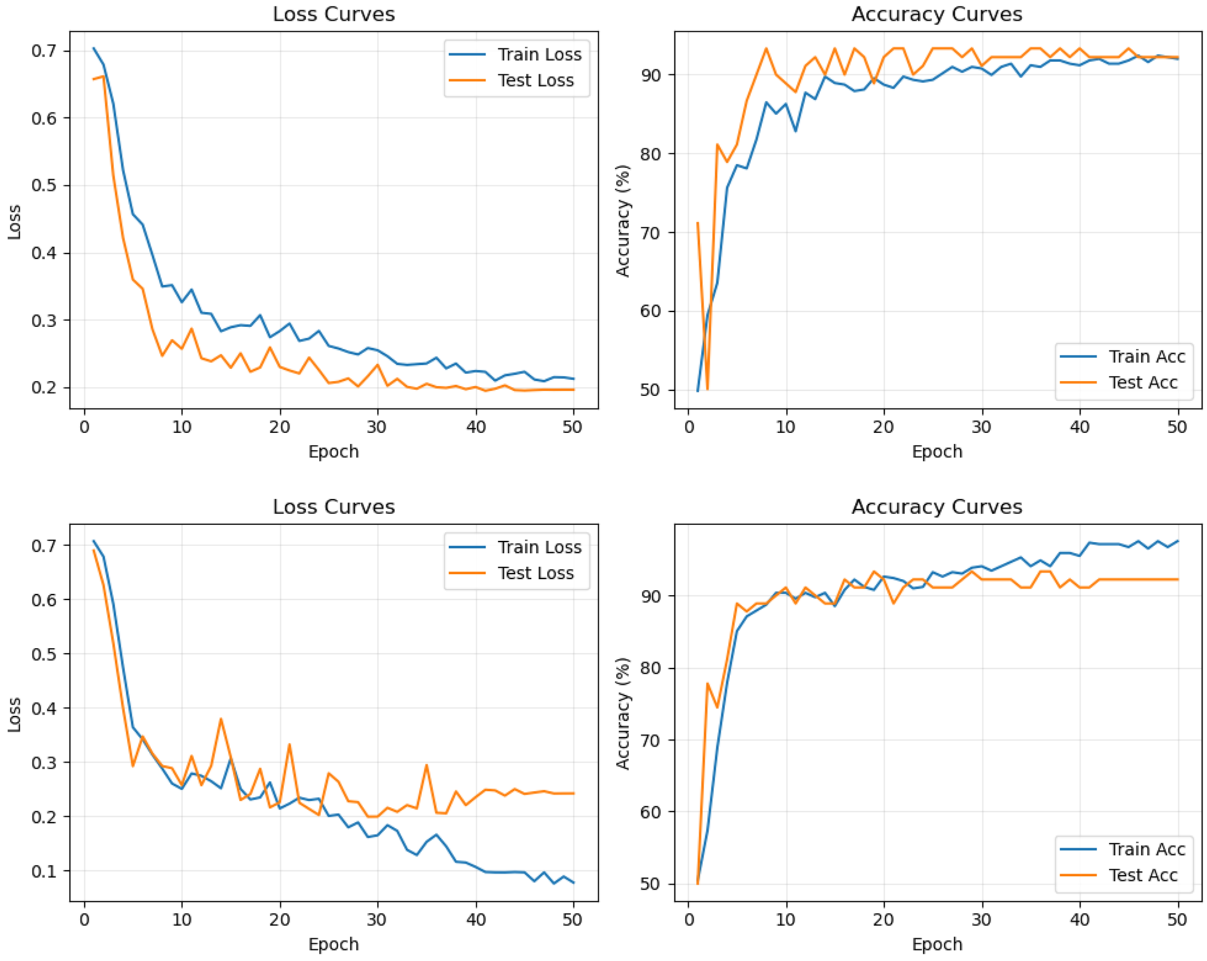

Figure 4. Accuracy for 50 epochs training and validation of BiLSTM--Attention model and Hybrid CNN--BiLSTM--Attention architecture. For short-term motion features that the recurrent encoder needs, the hybrid model is faster and stably converges by convolutional preprocessing. On the contrary, the BiLSTM-only model converges more slowly, and variances increase during the optimization process as its task is to more explicitly model both local and long-range temporal patterns directly from raw disparity trajectories.

6-6- Threshold Analysis and Sensitivity

In order to analyze how Temporal ASTRA behaves under various clinical decision criteria; a threshold sensitivity analysis was devised for the posterior risk scores of the model. Decision thresholds were then varied between 0.5 and

0.8 — a range often used for screening applications — and confusion matrices derived at every threshold, to assess variations in false-positive and false-negative rates. This analytical approach permits examination of how classification behaviour changes with the decision boundary change.

Using the ROC curve, the Youden index identified an optimal threshold of T = 0.778, corresponding to a sensitivity of 65.9% and a specificity of 99.2%. This operating point favors a balanced trade-off between false positives and false negatives and minimizes overall classification error. Such a setting is appropriate for contexts in which both missed cases and unnecessary referrals carry significant clinical or logistical cost.

Conversely, the first step in screening workflows is sensitive — this is particularly true in early-stage or community-based situations where the objective is to ensure no strabismus cases are overlooked. To account for this scenario, a high-sensitivity operating mode was assessed using a smaller decision threshold. A sensitivity increases to 95.5% and specificity decline to 21.2% was observed at T = 0.020. This difference is reflective of the projected rise in false positives for relaxed decision boundaries, which enable a greater number of borderline cases to be flagged.

Taken together, these results show that Temporal ASTRA does not operate at a single fixed point but can be adjusted to suit different screening objectives. The model can be configured either for balanced discrimination or for high-sensitivity pre-screening, depending on whether the emphasis is placed on minimizing missed cases or limiting referral burden. This flexibility is important for adapting the system to different clinical and deployment environments.

6-7- Results from Synthetic to Real Transfer

To examine how well Temporal ASTRA transfers from simulated vergence dynamics to real clinical recordings, the model was evaluated both with and without synthetic pretraining. As illustrated in Figure 5, the synthetic-to-real training strategy allows the network to first learn temporal vergence behavior under controlled conditions before being exposed to clinical data. When training was performed using only the available clinical videos, overall accuracy reached 82.0%, reflecting the difficulty of learning stable temporal representations from a small and heterogeneous dataset.

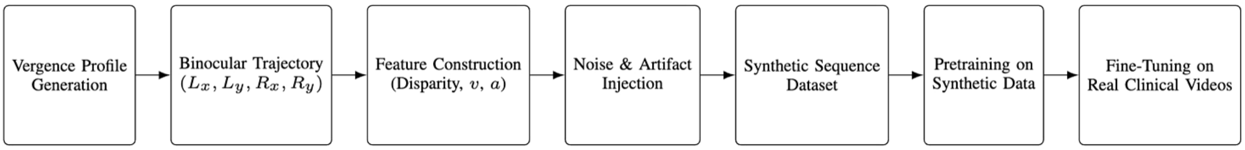
Vergence Profile Generation
Binocular Trajectory (L_x, L_y, R_x, R_y)
Feature Construction (Disparity, v, a)
Noise & Artifact Injection
Synthetic Sequence Dataset
Pretraining on Synthetic Data
Fine-Tuning on Real Clinical Videos

Figure 5. Top-to-bottom synthetic-to-real training pipeline for Temporal ASTRA. Synthetic vergence dynamics are used to generate binocular trajectories and motion features, followed by noise injection to approximate real-world signals. The synthetic dataset enables robust pretraining, which is later fine-tuned using real clinical video sequences.

When synthetic pretraining was introduced prior to fine-tuning on real clinical sequences, performance improved substantially. As reported in Table 4, clinical accuracy increased from 82.0% to 90.9%, accompanied by a corresponding improvement in AUC. This indicates that the model was better able to separate normal and abnormal temporal patterns after exposure to a wider range of vergence behaviors during pretraining.

The improvement can be attributed to the role of synthetic data in shaping the temporal representation space. The simulated sequences include controlled examples of drift, transient deviations, and noise, which expose the model to variations that may occur infrequently or inconsistently in clinical recordings. As a result, pretraining reduces overfitting to subject-specific or dataset-specific artifacts and improves generalization when the model is later fine-tuned on real patient data.

These findings highlight the practical value of synthetic data for temporal oculomotor analysis, particularly in clinical contexts where collecting large volumes of annotated video data is constrained by logistics, ethics, or patient availability.

Table 4. Effect of synthetic pretraining on real-world clinical video performance

Training Strategy	Accuracy	AUC
No Pretraining (Real Only)	82.0%	85.4%
Synthetic Pretrain + Fine-tune	**90.9%**	**93.7%**

6-8- Clinical Validation in the Real World

To test the robustness of Temporal ASTRA under practical conditions, a video-based evaluation was carried out using 24 real-world recordings obtained from the Dreamstime repository (12 normal and 12 strabismus cases; retrieved November 2025). The videos varied in subject age, lighting, camera distance, and eye movement behavior, and therefore provide a reasonable approximation of the variability expected in unconstrained screening settings.

Each video was processed through a fixed inference pipeline. Recordings were first divided into overlapping temporal windows of 5 seconds with 50% overlap. Pupil positions were extracted frame by frame using MediaPipe, and temporal features were constructed as described in Section 2. The fine-tuned Temporal ASTRA model produced a risk score for each window, and these window-level outputs were combined at the video level using majority voting together with mean risk aggregation. This resulted in a single risk score and binary classification for each video.

Per-video risk scores and corresponding ground-truth labels are reported in Table 5. At the operating threshold selected from the threshold sensitivity analysis, all strabismus cases were correctly identified, corresponding to 100% sensitivity on this dataset. Most normal videos produced low risk scores, although a small number were assigned higher values. These cases were often associated with transient tracking noise, large gaze shifts, or eye movement patterns that resembled brief instability rather than sustained misalignment.

Table 5. Video-level risk scores on the Dreamstime clinical dataset

Video ID	Clinical Label	Risk Score
138137050	Normal	0.0034
170400122	Normal	0.0000
175589778	Normal	0.0136
205252777	Normal	0.0065
227349987	Normal	0.8600
258201453	Normal	0.0848
258269600	Normal	0.1313
272049211	Normal	0.0000
357053905	Normal	0.2998
357054540	Normal	0.0000
366053856	Normal	0.0000
401644794	Normal	0.0000
124197414	Strabismus	1.0000
128812138	Strabismus	0.1387
142681115	Strabismus	1.0000
165165736	Strabismus	0.7161
205575906	Strabismus	0.0923
227177477	Strabismus	0.3231
249973691	Strabismus	1.0000
252652595	Strabismus	0.5577
276208579	Strabismus	0.2799
328289543	Strabismus	0.0256
372604642	Strabismus	0.4264
389486585	Strabismus	0.0599

Overall, these results show that the model is able to detect strabismus in real-world video recordings without requiring calibration. Some normal cases were still flagged as high risk, indicating that the choice of decision threshold has a direct impact on screening outcomes. This suggests that, while the approach is suitable for use with unconstrained video data, threshold selection must be adjusted according to the intended screening setting.

A threshold sweeps enabled consideration of an appropriate clinical decision boundary for real-world deployment. Since strabismus pre-screening prioritizes minimizing false negatives, we chose the threshold that offered the greatest sensitivity while maintaining reasonable specificity. For T = 0.020 the best operating value was found. At this intersection, the model reached the:

- Accuracy = 83.3%,
- Sensitivity = 100.0%,
- Specificity = 66.7%.

Using the optimal operating point obtained from the ROC and threshold–sensitivity sweep in Figure 6, the resulting classification behavior at T=0.020 is summarized in Table 6.

Table 6. Confusion matrix at the optimal pre-screening threshold (T = 0.020)

	Predicted Strabismus	Predicted Normal
Actual Strabismus	12 (TP)	0 (FN)
Actual Normal	4 (FP)	8 (TN)

Thus, these results demonstrate that *all* strabismus cases were indeed identified, with perfect sensitivity. Such behavior is desirable in a pre-screening environment, where the clinical goal is to ensure that no patients with potential ocular misalignment are missed, even at the expense of occasional false positives. The analysis of the four normal videos showed several common contributing factors, including (i) extreme gaze eccentricity, such that one eye approaches the field-of-view boundary; (ii) transient partial occlusions of the pupil and (iii) unstable fixation patterns closely resembling mild vergence drift. This introduces motion signatures that partially overlap with the original pathological pattern, and risk scores increase. However, these false alarms are clinically acceptable in early screening pipelines in which follow-up assessment is routine. Overall, the threshold sweep validates the ability of Temporal ASTRA to generalize to heterogeneous real-world video recordings despite the differences in subject appearance, lighting and recording conditions. Crucially, it has achieved this output without subject-specific calibration, highlighting its applicability to telemedicine and large-scale community screening projects.

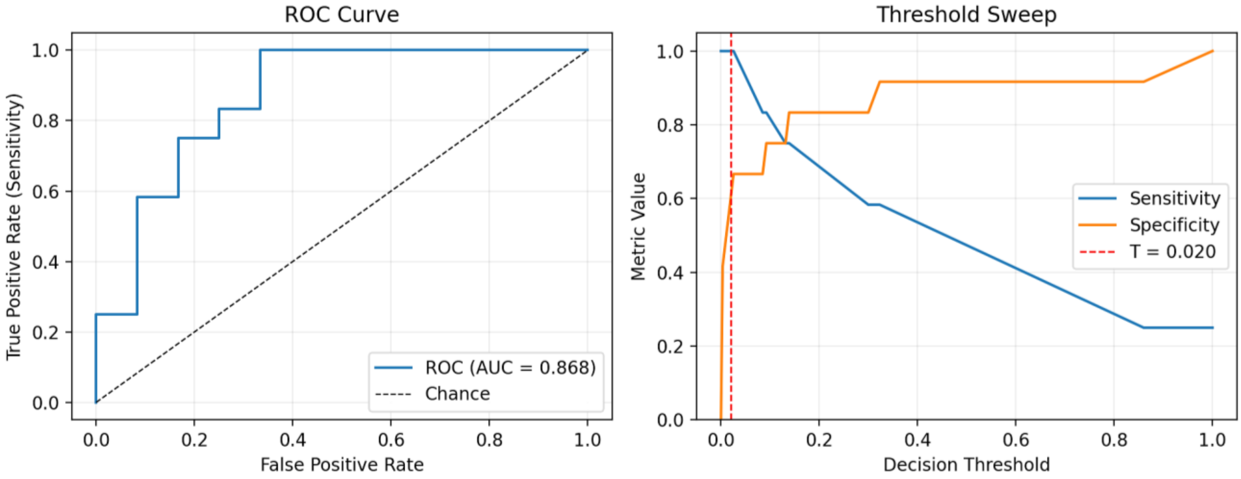

Figure 6. ROC and threshold–sensitivity sweep analysis of the Dreamstime clinical dataset. The ROC curve produced by video-level risk scores is presented on the left and includes the operating point at the screening threshold (T=0.020) marked in red. The right panel shows the sensitivity and specificity profiles across decision thresholds, demonstrating the trade-off between false negatives and false positives. The chosen threshold achieves 100% sensitivity, prioritizing safe pre-screening performance while tolerating a small number of false positives.

6-9- Temporal Stability and Sliding-Window Stability

Temporal ASTRA also presented robust temporal stability when applied to full clinical videos. Sliding window inference-based statistical approach together with temporal smoothing was used to obtain good prediction of sustained abnormal sequences, summarized in Table 7.

Many of the strabismus videos provided long, uninterrupted abnormal intervals that extended for 280–360 seconds, suggesting that the long-horizon vergence patterns that were simulated by the model were actually consistent with the measured behavior rather than being based solely on small aberrant frames. The observed observation shows temporal characteristics of the learned features are stable over large time bands, even under natural eye motion.

A subset of normal videos showed elevated and sometimes prolonged abnormal intervals (e.g., 70–110 seconds). From manual inspection, it was found that these correlated with gaze eccentricity, occlusion of one pupil or unstable fixations, all of which legitimately inflated the model's risk estimate under a pre-screening mode. Crucially, these disturbances did not induce rapid variation of the categories; predictions were consistent and temporally coherent.

Taken together, this results support that Temporal ASTRA exhibits clinically relevant temporal robustness. The framework retains resilience against temporary artifacts like blinks, micro-saccades, and transient tracking loss, but is sensitive to long-lived misalignment patterns which can persist over seconds to minutes of video. This temporal reliability is crucial for deployment in real world scenarios for continuous video-based strabismus pre-screening.

Table 7. Temporal consistency analysis on 24 clinical videos using sliding-window inference. Longest abnormal interval is computed at threshold T = 0.02

Video ID	Clinical Label	Average Risk	Longest (s)
138137050	Normal	0.0034	7.5
170400122	Normal	0.0000	0.0
175589778	Normal	0.0136	10.0
205252777	Normal	0.0065	27.5
227349987	Normal	0.8600	110.0
258201453	Normal	0.0848	25.0
258269600	Normal	0.1313	70.0
272049211	Normal	0.0000	0.0
357053905	Normal	0.2998	10.0
357054540	Normal	0.0000	0.0
366053856	Normal	0.0000	0.0
401644794	Normal	0.0000	0.0
124197414	Strabismus	1.0000	280.0
128812138	Strabismus	0.1387	20.0
142681115	Strabismus	1.0000	25.0
165165736	Strabismus	0.7161	325.0
205575906	Strabismus	0.0923	27.5
227177477	Strabismus	0.3231	335.0
249973691	Strabismus	1.0000	145.0
252652595	Strabismus	0.5577	367.5
276208579	Strabismus	0.2799	80.0
328289543	Strabismus	0.0256	22.5
372604642	Strabismus	0.4264	87.5
389486585	Strabismus	0.0599	25.0

6-10- Analysis for the Failure Model

A detailed review of misclassified videos revealed two main types of failure: (i) normal subjects assigned elevated risk scores (false positives) and (ii) strabismus subjects with relatively low but still positive risk estimates.

False positives (normal videos classified as strabismus)**:** Four normal videos produced high risk scores, with the highest observed in video 227349987 (risk = 0.8600). Visual inspection of these recordings showed several factors that are known to interfere with vergence-based temporal analysis. These included prolonged gaze excursions, partial eyelid occlusion, short periods of reduced pupil visibility, and moderate head motion. Each of these factors can introduce brief spikes in estimated interocular disparity that resemble transient divergence events. Because Temporal ASTRA is designed to favor sensitivity, such short-lived fluctuations can accumulate during temporal smoothing and lead to a positive classification. In the context of pre-screening, this behavior reflects a deliberate bias toward minimizing missed cases rather than an outright system failure, although it increases false-positive rates.

Low-risk strabismus cases**:** Several strabismus videos (e.g., 389486585, 205575906, and 328289543) produced relatively low risk scores, ranging from 0.026 to 0.093. These cases were characterized by mild intermittent exotropia, with short divergence episodes of small amplitude that appeared only briefly within otherwise stable fixation periods. Although these cases were correctly identified at the high-sensitivity threshold (T = 0.020), the results indicate that they would be at risk of misclassification if a more conservative threshold were used. This highlights the dependence of detection performance on threshold selection, particularly for subtle or early-stage conditions.

Primary failure pattern: The most diagnostically challenging example was video 165165736 (risk = 0.7161); it had short bursts of divergence embedded within long periods of stable alignment. Although correctly classified, the temporal profile of this case illustrates the limitations of the current aggregation strategy. When brief misalignment events are surrounded by long segments of normal behavior, temporal smoothing can reduce their influence on the final risk estimate. As a result, subtle but clinically relevant deviations may be underrepresented unless additional mechanisms are introduced to emphasize short-duration events.

Taken together, these failure modes suggest that the dominant source of error arises not from missing clear strabismus cases, but from sensitivity to normal gaze behavior that transiently resembles pathological divergence. This behavior is consistent with a screening system optimized for high sensitivity. At the same time, it indicates directions for future improvement, such as adaptive temporal weighting, multi-scale aggregation, or threshold adjustment, to improve specificity without compromising clinical safety.

6-11- Calibration

To determine whether Temporal ASTRA produces probabilistic estimates relevant to actual risk estimates in the clinic, we tested calibration quality in terms of reliability plots, Brier scores, and ECE. Prior to calibration, the model was shown to over-predict the probability distributions using an ECE of 0.177 which suggested a noticeable mismatch between predicted probabilities and the actual empirical frequencies of strabismus events.

The temperature scaling was then used as a post-hoc calibration, and an optimal temperature parameter of T = 2.07 was set. After scaling, model outputs displayed significantly stronger correlation with the observed accuracy of each probability bin, leading to an ECE of 0.145. This enhancement indicates improved mapping between model-estimated strabismus risk and the true likelihood of pathology. Although the numerical shift of the calibration might not seem drastic, we find the clinical results substantial. Calibration of probabilities in screening contexts increases interpretability for clinicians but also allows for threshold selection that better reflects desired trade-offs in sensitivity–specificity. Additionally, calibrated risk estimates enable downstream applications, including risk stratification, telemedicine triage, and automated referral decisions. Overall, the calibration data verified that Temporal ASTRA can be implemented not only as a high-sensitivity classifier but also as a reliable probabilistic estimator of binocular misalignment.

7- Discussion

This study shows that temporal information is important for strabismus pre-screening. Instead of judging eye alignment from a single image, the proposed method examines how binocular coordination changes over several seconds. This matches how clinicians typically assess patients, by observing fixation stability, slow drift, or brief loss of fusion during sustained viewing rather than relying on a single moment. The results show that abnormal cases often differ from normal ones in how alignment evolves over time.

The experiments also show that screening without calibration is possible when relative motion is used instead of absolute gaze position. Tracking changes in interocular disparity over time reduces sensitivity to anatomical differences and recording conditions that usually affect uncalibrated gaze estimation. This is especially relevant in pediatric and telemedicine settings, where controlled environments and subject-specific calibration are difficult to achieve.

Results from unseen subjects indicate that abnormal vergence tends to follow similar temporal patterns across individuals. This consistency is important for large-scale screening because it suggests that the system does not need to be adjusted for each subject. Such behavior is suitable for community or school-based screening, where the process needs to be simple and fast.

The improvement obtained with synthetic pretraining shows the value of simulation when clinical data are limited. The synthetic sequences expose the model to vergence behaviors that may appear rarely or inconsistently in real recordings. In this way, simulation helps the model learn a broader range of temporal patterns rather than simply increasing the amount of training data.

Several limitations remain. The simulator does not include some real-world factors, such as eyelid motion, strong lighting changes, or large head movements, which can affect pupil tracking. In addition, very brief or subtle deviation episodes may be reduced by temporal averaging. These effects can lead to false positives or reduced sensitivity in borderline cases.

Future work should focus on improving robustness under challenging recording conditions and expanding the simulated behaviors to better reflect clinical variability. From a modeling perspective, alternative temporal encoders or adaptive windowing may help capture both short events and longer trends. Further clinical evaluations, particularly in pediatric and telemedicine settings, will be needed to assess usability and screening impact.

Overall, analyzing binocular coordination over time provides a practical basis for calibration-free strabismus pre-screening. Focusing on dynamic behavior rather than static alignment is closer to clinical assessment and supports the use of video-based screening in unconstrained environments.

7-1- Comparison with Previous Studies

Automated strabismus screening methods generally fall into two groups: static image-based approaches and calibrated gaze-tracking systems. Static photoscreening methods, including Hirschberg-based techniques and recent deep learning classifiers applied to single images, have reported good performance for detecting large, constant deviations under controlled conditions [3–5, 9]. However, because these methods rely on a single image, they are limited in detecting intermittent exotropia, vergence instability, or fatigue-related deviations that may not be present at the time of capture [4, 6, 18].

Calibrated eye-tracking and gaze estimation systems have shown high precision in laboratory and clinical environments, particularly for studying fixation stability, saccades, and smooth pursuit [9, 10, 24, 25]. Several studies have added temporal modeling to better capture oculomotor behavior [8, 11, 27, 24]. Despite this, most gaze-based systems require per-subject calibration and controlled viewing geometry, which limits their use in pediatric screening and telemedicine settings [35–39].

Temporal ASTRA differs from these approaches by not estimating absolute gaze direction. Instead, it focuses on relative vergence dynamics, which reduces sensitivity to anatomical variation and camera setup while retaining clinically relevant temporal information. Prior work using recurrent and attention-based models has shown that temporal representations improve performance in gaze and oculomotor analysis [8, 11, 24, 25]. The present study builds on this work by targeting calibration-free strabismus pre-screening rather than precise gaze estimation.

In quantitative terms, the performance of Temporal ASTRA is comparable to, and in some cases higher than, that reported by static deep learning models used for strabismus screening, while maintaining high sensitivity under unconstrained recording conditions [4, 5, 21]. This performance is achieved without controlled illumination, fixed viewing distance, or subject-specific calibration. The observed sensitivity–specificity trade-offs are consistent with those reported in screening-focused studies and reflect a preference for sensitivity in early detection [3, 21].

Overall, these comparisons suggest that Temporal ASTRA complements existing screening and diagnostic tools by addressing situations where static or calibration-dependent methods are less effective. Rather than replacing clinical examinations, the framework is intended to function as a pre-screening step that captures dynamic misalignment patterns and supports scalable use in community and telemedicine settings.

8- Conclusion

This work brought Temporal ASTRA, a fully calibration-free temporal modeling framework for automated strabismus pre-screening. The system is implemented with four essential algorithmic elements: (i) a synthetic dynamic vergence generation approach to allow for large-scale pretraining, (ii) a hybrid CNN--BiLSTM that captures short-term oculomotor fluctuations and long-range temporal dependencies, (iii) an attention-based temporal pooling mechanism that adaptively emphasizes diagnostically informative parts of the sample, and (iv) a sliding-window inference and aggregation approach to stabilize predictions under real-world noise scenarios. Collectively, these components allow reliable estimation of binocular alignment dynamics directly from raw pupil trajectories, with no need for geometric calibration per subject.

Full-blown experiments reveal that the temporal progression of interocular disparity offers a much stronger biomarker of misalignment than static frame-specific appearance cues. Ablation analyses revealed the largest performance drivers to be temporal recurrence and sequence normalization, while synthetic-to-real pretraining improved out-of-distribution generalization significantly. Real-time clinical investigation also revealed that Temporal ASTRA holds high sensitivity under diverse acquisition situations in hardware and outputs calibrated probabilities after scaling the temperature, supporting it for application for downstream decision-support pipelines.

Although these developments have been made, still the system has the challenge of ultra-short discontinuities on an intermittent basis for which the temporal smoothing can cause disturbance. Future studies will involve multi-scale temporal encoders, transformer-based sequence models, and probabilistic uncertainty quantification to improve fine-grained temporal sensitivity.

Extending the synthetic engine to a larger range of physiological features, including micro-saccadic noise and latent phoria decompensation, will improve domain transfer further. The proposed Temporal ASTRA framework acts as a technically solid platform for achieving high fidelity in scalable, calibration-free, temporal-aware strabismus detection that can be employed in both telemedicine settings as well as on-line monitor systems.

9- Declarations

9-1- Author Contributions

Conceptualization, L.U. and W.K.; methodology, W.K.; software, W.K.; validation, L.U., W.K., and P.K.; formal analysis, W.K.; investigation, L.U.; resources, W.K.; data curation, P.K.; writing—original draft preparation, L.U. and W.K.; writing—review and editing, W.K.; visualization, W.K.; supervision, L.U.; project administration, W.K.; funding acquisition, W.K. All authors have read and agreed to the published version of the manuscript.

9-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

9-3- Funding

Support provided by Walailak University and the Office of the National Digital Economy and Society Commission (ONDE) is gratefully acknowledged, as it was instrumental to the completion of this research. This project was also carried out under the Reinventing University program, supported by the Ministry of Higher Education, Science, Research and Innovation.

9-4- Institutional Review Board Statement

The research protocol titled 'Intelligent System for Gaze Analysis and Ocular Abnormality Screening Using Deep Learning Techniques on Secondary Datasets' (Project Code: WU-EC-EN-4-461-68) was reviewed by the Human Research Ethics Committee of Walailak University. On January 15, 2026, the committee determined that the study qualified for an exemption from full ethical review, as the methodology involves the use of secondary datasets. This approval is documented under Certification No. WUEC-26-020-01.

9-5- Informed Consent Statement

Not applicable.

9-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

9-7- Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the authors utilized Google Gemini for language polishing and structural refinement. Following the use of this tool, the content was thoroughly reviewed and edited by the authors to ensure technical accuracy and alignment with the research objectives. The authors take full responsibility for the integrity of the published work.

10- References

[1] Noer, M. H. G., Prastyani, R., Fahmi, A., & Loebis, R. (2024). Strabismus and Binocular Vision: A Comprehensive Review of Pathophysiology, Risk Factors, Classification, Diagnostic, and Treatment. International Journal of Scientific Advances, 5(6), 1532–1536. doi:10.51542/ijscia.v5i6.80.

[2] Sawamura, H., Gillebert, C. R., Todd, J. T., & Orban, G. A. (2018). Binocular stereo acuity affects monocular three-dimensional shape perception in patients with strabismus. British Journal of Ophthalmology, 102(10), 1413–1418. doi:10.1136/bjophthalmol-2017-311393.

[3] Silbert, D.I., Matta, N.S., Chang, L. (2025). Photoscreening. EyeWiki, American Academy of Ophthalmology, San Francisco, United States. Available online: https://eyewiki.aao.org/Photoscreening (accessed on January 2026).

[4] Hartness, E. M., Jiang, F., Zamba, G. K. D., Allen, C., Bragg, T. L., Nellis, J., Dumitrescu, A. V., & Kardon, R. H. (2025). Automated strabismus evaluation: a critical review and meta-analysis. Frontiers in Neurology, 16, 1620568. doi:10.3389/fneur.2025.1620568.

[5] Karaaslan, Ş., Kobat, S. G., & Gedikpınar, M. (2023). A new method based on deep learning and image processing for detection of strabismus with the Hirschberg test. Photodiagnosis and Photodynamic Therapy, 44, 103479. doi:10.1016/j.pdpdt.2023.103805.

[6] Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. Behavior Research Methods, 42(3), 692–700. doi:10.3758/BRM.42.3.692.

[7] Hess, R. F., & Thompson, B. (2015). Amblyopia and the binocular approach to its therapy. Vision Research, 114, 4–16. doi:10.1016/j.visres.2015.02.009.

[8] Zheng, Y., Fu, H., Li, R., Lo, W. L., Chi, Z., Feng, D. D., Song, Z., & Wen, D. (2019). Intelligent evaluation of strabismus in videos based on an automated cover test. Applied Sciences (Switzerland), 9(4), 731. doi:10.3390/app9040731.

[9] Yarkheir, M., Sadeghi, M., Azarnoush, H., Akbari, M. R., & Khalili Pour, E. (2025). Automated strabismus detection and classification using deep learning analysis of facial images. Scientific Reports, 15(1), 3910. doi:10.1038/s41598-025-88154-6.

[10] Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2019). MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(1), 162–175. doi:10.1109/TPAMI.2017.2778103.

[11] Hashemi, H., Pakzad, R., Heydarian, S., Yekta, A., Aghamirsalim, M., Shokrollahzadeh, F., Khoshhal, F., Pakbin, M., Ramin, S., & Khabazkhoob, M. (2019). Global and regional prevalence of strabismus: a comprehensive systematic review and meta-analysis. Strabismus, 27(2), 54–65. doi:10.1080/09273972.2019.1604773.

[12] Silva, N., Castro, C., Caiado, F., Maia, S., Miranda, V., Parreira, R., & Menéres, P. (2022). Evaluation of Functional Vision and Eye-Related Quality of Life in Children with Strabismus. Clinical Ophthalmology, 16, 803–813. doi:10.2147/OPTH.S354835.

[13] Cotter, S. A., Varma, R., Tarczy-Hornoch, K., McKean-Cowdin, R., Lin, J., Wen, G., Wei, J., Borchert, M., Azen, S. P., Torres, M., Tielsch, J. M., Friedman, D. S., Repka, M. X., Katz, J., Ibironke, J., & Giordano, L. (2011). Risk factors associated with childhood strabismus: The multi-ethnic pediatric eye disease and Baltimore pediatric eye disease studies. Ophthalmology, 118(11), 2251–2261. doi:10.1016/j.ophtha.2011.06.032.

[14] Khazaei, S., & Mansori, K. (2018). Variability of preoperative measurements in intermittent exotropia and its effect on surgical outcome. Journal of American Association for Pediatric Ophthalmology and Strabismus, 22(4), 332. doi:10.1016/j.jaapos.2017.06.027.

[15] Hatt, S. R., Leske, D. A., Liebermann, L., & Holmes, J. M. (2015). Quantifying variability in the measurement of control in intermittent exotropia. Journal of AAPOS, 19(1), 33–37. doi:10.1016/j.jaapos.2014.10.017.

[16] Tengtrisorn, S., Tungsattayathitthan, A., Na Phatthalung, S., Singha, P., Rattanalert, N., Bhurachokviwat, S., & Chouyjan, S. (2021). The reliability of the angle of deviation measurement from the Photo-Hirschberg tests and Krimsky tests. PLoS ONE, 16(12 December), 258744. doi:10.1371/journal.pone.0258744.

[17] Zheng, C., Yao, Q., Lu, J., Xie, X., Lin, S., Wang, Z., Wang, S., Fan, Z., & Qiao, T. (2021). Detection of referable horizontal strabismus in children's primary gaze photographs using deep learning. Translational Vision Science and Technology, 10(1), 1–9. doi:10.1167/tvst.10.1.33.

[18] Economides, J. R., Adams, D. L., & Horton, J. C. (2016). Variability of ocular deviation in strabismus. JAMA Ophthalmology, 134(1), 63–69. doi:10.1001/jamaophthalmol.2015.4486.

[19] Valenti, R., Sebe, N., & Gevers, T. (2012). Combining head pose and eye location information for gaze estimation. IEEE Transactions on Image Processing, 21(2), 802–815. doi:10.1109/TIP.2011.2162740.

[20] Wu, D., Li, Y., Zhang, H., Yang, X., Mao, Y., Chen, B., Feng, Y., Chen, L., Zou, X., Nie, Y., Yin, T., Yang, Z., Liu, J., Shang, W., Yang, G., & Liu, L. (2024). An artificial intelligence platform for the screening and managing of strabismus. Eye (Basingstoke), 38(16), 3101–3107. doi:10.1038/s41433-024-03228-5.

[21] de Araújo Santos, R. D., de Almeida, J. D. S., Teixeira, J. A. M., Valente, T. L. A., Braz, G., & de Paiva, A. C. (2025). Automated strabismus diagnosis: A deep learning approach using cover test video analysis. Engineering Applications of Artificial Intelligence, 159. doi:10.1016/j.engappai.2025.111161.

[22] Wagle, N., Morkos, J., Liu, J., Reith, H., Greenstein, J., Gong, K., Gangan, I., Pakhomov, D., Hira, S., Komogortsev, O. V., Newman-Toker, D. E., Winslow, R., Zee, D. S., Otero-Millan, J., & Green, K. E. (2022). aEYE: A deep learning system for video nystagmus detection. Frontiers in Neurology, 13, 963968. doi:10.3389/fneur.2022.963968.

[23] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[24] Zemblys, R., Niehorster, D. C., & Holmqvist, K. (2019). gazeNet: End-to-end eye-movement event detection with deep neural networks. Behavior Research Methods, 51(2), 840–864. doi:10.3758/s13428-018-1133-5.

[25] Krakowczyk, D. G., Prasse, P., Reich, D. R., Lapuschkin, S., Scheffer, T., & Jäger, L. A. (2023). Bridging the Gap: Gaze Events as Interpretable Concepts to Explain Deep Neural Sequence Models. Eye Tracking Research and Applications Symposium (ETRA), 3588412. doi:10.1145/3588015.3588412.

[26] Cheng, Y., Wang, H., Bao, Y., & Lu, F. (2024). Appearance-Based Gaze Estimation with Deep Learning: A Review and Benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7509–7528. doi:10.1109/TPAMI.2024.3393571.

[27] Mokatren, M., Kuflik, T., & Shimshoni, I. (2024). Calibration-Free Mobile Eye-Tracking Using Corneal Imaging. Sensors, 24(4), 1237. doi:10.3390/s24041237.

[28] Wood, E., Baltruaitis, T., Zhang, X., Sugano, Y., Robinson, P., & Bulling, A. (2015). Rendering of eyes for eye-shape registration and gaze estimation. Proceedings of the IEEE International Conference on Computer Vision, 2015 International Conference on Computer Vision, ICCV 2015, 3756–3764. doi:10.1109/ICCV.2015.428.

[29] Garde, G., Larumbe-Bergera, A., Bossavit, B., Cabeza, R., Porta, S., & Villanueva, A. (2020). Gaze estimation problem tackled through synthetic images. Eye Tracking Research and Applications Symposium (ETRA), 1–5. doi:10.1145/3379156.3391368.

[30] Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-January, 2242–2251. doi:10.1109/CVPR.2017.241.

[31] Fuhl, W., Kuebler, T., Brinkmann, H., Rosenberg, R., Rosenstiel, W., & Kasneci, E. (2018). Region of interest generation algorithms for eye tracking data. Proceedings of the 3rd Workshop on Eye Tracking and Visualization, 1–9. doi:10.1145/3205929.3205937.

[32] Chen, Z., Fu, H., Lo, W. L., & Chi, Z. (2018). Strabismus Recognition Using Eye-Tracking Data and Convolutional Neural Networks. Journal of Healthcare Engineering, 7692198. doi:10.1155/2018/7692198.

[33] Yan, Z., Wu, Y., Shan, Y., Chen, W., & Li, X. (2022). A dataset of eye gaze images for calibration-free eye tracking augmented reality headset. Scientific Data, 9(1), 115. doi:10.1038/s41597-022-01200-0.

[34] Wood, E., Baltrušaitis, T., Morency, L. P., Robinson, P., & Bulling, A. (2016). Learning an appearance-based gaze estimator from one million synthesised images. Eye Tracking Research and Applications Symposium (ETRA), 14, 131–138. doi:10.1145/2857491.2857492.

[35] Duchowski, A. T. (2017). Eye tracking methodology: Theory and practice: Third edition. In Eye Tracking Methodology: Theory and Practice: Third Edition. Springer, Cham, Switzerland. doi:10.1007/978-3-319-57883-5.

[36] Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye Tracking for Everyone. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2176–2184. doi:10.1109/CVPR.2016.239.

[37] Alnajar, F., Gevers, T., Valenti, R., & Ghebreab, S. (2013). Calibration-free gaze estimation using human gaze patterns. Proceedings of the IEEE International Conference on Computer Vision, 137–144. doi:10.1109/ICCV.2013.24.

[38] Jin, S., Dai, J., & Nguyen, T. (2023). Kappa Angle Regression with Ocular Counter-Rolling Awareness for Gaze Estimation. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2023-June, 2659–2668. doi:10.1109/CVPRW59228.2023.00266.

[39] Zhao, Z., Meng, H., Li, S., Wang, S., Wang, J., & Gao, S. (2025). High-Accuracy Intermittent Strabismus Screening via Wearable Eye-Tracking and AI-Enhanced Ocular Feature Analysis. Biosensors, 15(2), 110. doi:10.3390/bios15020110.