Universiteit Leiden

Master Computer Science

EveLoad: An Event-Based Eye-Tracking Dataset with Gaze Annotations and Cognitive Load Labels

Name:	Shaohua Guan
Student ID:	s3777626
Date:	29/01/2026
Specialisation:	Advanced Computing and Systems
1st supervisor:	Qinyu Chen
2nd supervisor:	Zhaochun Ren

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Contents

Acknowledgements

I would like to sincerely thank my supervisor, Dr. Qinyu Chen, for her detailed guidance, patience, and continuous support throughout this work. Her professional mentorship and respectful, open attitude made it easy to discuss ideas and work through problems collaboratively.

I would also like to thank my co-supervisor, Dr. Zhaochun Ren, for his constructive suggestions and for raising important considerations regarding the practical challenges this project may face in real-world applications.

I am grateful to Dr. Guorui Lu for proposing the idea of incorporating cognitive workload into the study, which played an important role in shaping the direction of this research.

I also thank my boyfriend, Pu Li, and my friend, Shan Jiang, for their care and support during my recovery from a fracture while working on this thesis.

Chapter 1

Introduction

In the pursuit of heightened immersion and natural interaction, virtual reality (VR) systems face stringent requirements for high visual fidelity and low motion-to-photon latency [4]. VR systems must continuously deliver stereoscopic graphics with high resolution, a wide field of view, and low latency. Any additional latency can diminish the perceived realism of interaction and potentially cause discomfort. Concurrently, limitations in size, weight, power consumption, and thermal management hinder head-mounted devices from relying solely on hardware scaling to meet the increasing demands of rendering and interaction. Consequently, achieving high-quality and low-latency experiences remains a significant challenge in VR system design.

Perception-driven spatial optimization offers an effective approach to this challenge. In particular, foveated rendering leverages the human visual system's high acuity in central vision and reduced acuity in peripheral vision to allocate more computation to perceptually important regions [11], thereby preserving perceived quality while reducing rendering cost [4]. Moreover, gaze-contingent foveated rendering relies on real-time eye tracking to locate the gaze point and align the high-quality region accordingly, improving perceptual efficiency [8]. Therefore, developing a low-latency and reliable gaze-sensing pipeline is critical for dynamic foveated rendering under wearable constraints.

State-of-the-art gaze sensing for VR typically relies on dedicated eye trackers or frame-based RGB cameras [10]. However, in head-mounted settings, rapid eye movements, close-range imaging, and constrained or unstable illumination conditions can degrade frame-based imaging quality, introducing additional sensing and processing delay while reducing robustness. These factors limit the reliability of low-latency gaze-

point updates required for gaze-contingent rendering [1]. To address these challenges, an event camera reports pixel-level brightness changes asynchronously, offering high temporal resolution and wide dynamic range. By avoiding motion blur under rapid motion and reducing redundant data capture, event-based sensing provides a promising modality for low-latency and power-efficient eye tracking in wearable systems.

Cognitive-science studies further suggest that changes in mental workload can affect attentional resource allocation and may lead to a shrinkage of the useful field of view (UFOV), accompanied by reduced sensitivity to peripheral information [6] [2]. This implies that the optimal spatial extent of the high-quality region in foveated rendering should not be treated as constant, but may vary with the user's cognitive state. Meanwhile, existing event-based gaze-estimation studies primarily focus on geometric accuracy and latency, and typically do not include mental-workload labels as an explicit cognitive-state variable [17]. This omission makes it difficult to systematically analyze the interaction between gaze behavior and mental workload under controlled and reproducible conditions.

These observations raise a key question: is whether, within an event-based eye sensing framework, it is possible to simultaneously and reliably estimate gaze location and cognitive load, and to systematically characterize the attainable performance limits and failure boundaries, thereby providing a foundation for cognitively adaptive interaction and rendering strategies.

For mental-workload recognition, most prior eye-tracking studies adopt free-viewing or central-stimulus paradigms [13]. In such settings, workload differences often manifest in behavioral cues such as the spatial dispersion of gaze trajectories, overall area coverage, and large saccades, which provide strong discriminative signals for machine-learning models. However, in task-driven interactive scenarios, gaze behavior is often continuously guided by explicit targets, substantially reducing the spatial freedom of gaze positions. Under such constraints, reliance on spatial-distribution cues becomes less effective, and workload recognition may need to depend more on fine-grained ocular dynamics such as microsaccades and blinks. This issue is especially pronounced under the target-guided conditions considered in this study, where gaze trajectories are intentionally constrained by task demands.

In light of the above motivations, we construct and release an event-camera-based eye-tracking dataset for task-driven gaze behavior. The dataset includes eye event streams, gaze-point coordinates, and mental-workload labels graded using an N-back task, a widely adopted and validated paradigm for controlled manipulation of working-memory load [5]. To better reflect interactive VR scenarios and reduce reliance on "free

exploration" cues such as gaze spatial dispersion for workload discrimination, we adopt a full-screen, randomly guided target-following paradigm that enforces task-driven and spatially constrained gaze behavior.

On this basis, we define two benchmark tasks—gaze prediction and cognitive load classification—and establish strong baseline models to systematically evaluate the learnability of memory load from event-based eye signals, as well as the achievable prediction accuracy.

In summary, this paper establishes a reproducible benchmark for event-based eye-tracking research under task-driven and gaze-constrained settings. It evaluates the achievable performance and failure thresholds for gaze-point estimation and mental-workload recognition, and discusses implications for future cognitively adaptive interaction and rendering strategies.

The primary contributions of this paper are as follows:

- **An event-based eye-tracking dataset with cognitive load labels.** We construct and release an event-based eye-tracking dataset that provides eye event streams, synchronized RGB eye videos, screen-space gaze coordinates, and graded cognitive load labels based on the N-back task. To the best of our knowledge, this is the first event-based gaze dataset that jointly annotates fixation behavior and cognitive load, with the explicit goal of enabling cognitively adaptive gaze-contingent rendering beyond pure gaze localization.

- **A task-driven, gaze-constrained paradigm and benchmarking protocol.** We propose a full-screen random target-guided tracking paradigm that enforces task-driven and spatially constrained gaze behavior. Based on this design, we establish benchmark protocols and data split schemes for both gaze prediction and cognitive load classification, providing a reproducible evaluation framework aligned with interactive VR settings.

- **Baseline models and empirical validation.** We implement and evaluate strong baseline models, demonstrating the learnability of memory load from event-based eye signals.

Chapter 2

Related Work

This section reviews prior work related to our study from a system-level perspective and, based on this review, clarifies the positioning and research gaps addressed by this paper. Our organization is motivated by a key observation from cognitive science: variations in memory and mental workload influence the allocation of attentional resources, potentially leading to a contraction of the Useful Field of View (UFOV) and reduced sensitivity to peripheral information [2] [6] [3] [4]. This suggests that, in perceptually driven rendering systems, the optimal spatial extent of high-quality regions and the corresponding resource allocation strategies may need to be dynamically adjusted according to the user's cognitive load.

Motivated by this observation, we first review research on eye-movement-based mental workload estimation, summarizing commonly used workload manipulation and labeling paradigms (e.g., the N-back task) as well as eye-movement cues associated with workload changes. Existing studies typically focus on gaze localization and lack joint annotation and systematic evaluation of cognitive states such as mental workload within a unified data framework. Building on this observation, we further examine how cognitive state may influence perception-driven rendering and interaction strategies, and argue that jointly estimating gaze and cognitive load within a unified event-based eye-sensing pipeline is of fundamental importance for constructing cognitively adaptive systems. Finally, we discuss the relationship between cognitive state and perceptually driven rendering and interaction strategies, highlighting the potential significance of jointly sensing gaze and mental workload within a unified perceptual pipeline.

Through this review, we establish the background for the dataset construction, task definitions, and experimental evaluation presented in this paper, and motivate

our study of a unified benchmark spanning event-based eye sensing, gaze estimation, and mental workload.

2.1 Eye-Movement-Based Mental Workload Estimation

Mental workload is a core variable in human factors and interactive system research, reflecting the amount of cognitive resources consumed by an individual while performing a task. It is closely related to task performance, attentional allocation, and subjective experience [16, 14]. In experimental studies, the N-back task has been widely adopted as a controllable method for inducing graded levels of mental workload [5]. By varying the memory span (e.g., 0-back, 1-back, 2-back), different levels of working memory demand can be elicited under relatively stable experimental conditions, providing reliable workload labels for modeling [5]. Due to its high controllability and reproducibility, the N-back paradigm has been used not only in eye-tracking studies but also in multimodal workload assessment involving EEG, heart rate variability, electrodermal activity, and related signals.

A substantial body of work has investigated whether eye-movement signals can reflect mental workload. Prior studies have identified various workload-related ocular cues, including blink rate and blink duration, microsaccades, saccade and fixation statistics, fixation stability, and—when measurable—pupil diameter variations [13, 3]. These features capture changes in cognitive resource allocation to some extent and have been used to build mental workload recognition models, ranging from traditional feature-engineering-based classifiers to more recent end-to-end learning approaches.

However, it is important to note that many existing eye-movement-based workload studies adopt experimental paradigms such as free viewing or central-stimulus setups [13]. In these settings, stimuli are typically concentrated in the center or a limited region of the display, and participants' gaze trajectories exhibit strong spatial separability. As a result, models can often rely on the spatial distribution of gaze trajectories, coverage patterns, or large-amplitude saccadic behavior to distinguish workload levels [3]. In contrast, in task-driven interactive scenarios, users' gaze is often continuously guided by task goals, leading to reduced spatial freedom of fixation. Under such constraints, the effectiveness of workload discrimination based primarily on gaze trajectory geometry may diminish, and finer-grained dynamic signals such as microsaccades and blinking behavior may become more critical [13].

At the sensing level, most existing eye-movement-based workload studies rely on frame-based RGB cameras, infrared cameras, or conventional eye trackers [10]. Although these systems provide reliable fixation and saccade statistics under controlled laboratory settings, their frame-based sampling mechanisms inherently limit temporal resolution. Subtle and rapid eye dynamics, such as microsaccades or transient blinking patterns that may encode fine-grained cognitive fluctuations, can be undersampled or temporally blurred [1]. Consequently, capturing high–temporal-resolution ocular dynamics under wearable and interactive constraints remains challenging, which may restrict the sensitivity and robustness of mental workload estimation models.

2.2 Event-Based Gaze Tracking

Event cameras output asynchronous, pixel-level brightness changes with high temporal resolution and a wide dynamic range, and they naturally avoid motion blur commonly observed in frame-based imaging during rapid movements [7]. Leveraging these properties, recent years have seen a growing body of work on event-camera-based eye movement and gaze estimation [1, 17]. These studies primarily focus on how to construct learnable representations from event streams—such as event voxel grids, time surfaces, event frames, or statistical features—and how to regress or classify gaze positions based on these representations to achieve low-latency or high-frequency gaze sensing [1, 17].

Although prior work has demonstrated that event-based eye signals are learnable for gaze estimation, most existing studies formulate the problem as a gaze-only perception task, with gaze position as the sole output, and do not incorporate annotations or analyses related to cognitive state [1, 17]. As a result, these methods are unable to address questions that are critical for interactive systems, such as whether eye dynamics and gaze estimation reliability change systematically under different levels of mental workload. This limitation constrains the applicability of event-based gaze sensing in cognitively adaptive interaction and rendering systems.

To address this gap, our work combines event-based gaze tracking with explicit mental workload annotation, enabling both gaze estimation and workload recognition within a unified dataset and experimental paradigm. To the best of our knowledge, the dataset constructed and released in this paper is the first gaze tracking dataset that simultaneously provides event-based eye data, ground-truth gaze coordinates, and graded mental workload labels based on the N-back task. This dataset establishes a reproducible benchmark for future studies and facilitates systematic comparison across

methods.

2.3 Cognitive State and Rendering: From Gaze to Fine-Grained Perceptual Strategies

Foveated rendering exploits the characteristics of the human visual system, which exhibits high acuity in central vision and reduced sensitivity in the periphery, by allocating computational resources preferentially to regions around the gaze point. This approach can significantly reduce rendering cost while maintaining perceived visual quality [4]. In dynamic foveated rendering, the high-quality region moves in real time with the user's gaze, making end-to-end latency and stability of gaze sensing critical factors for rendering quality. Excessive tracking latency can lead to misalignment between the high-quality region and the true gaze position, introducing perceptible artifacts and degrading user experience [9]. Consequently, low-latency and robust gaze sensing is a fundamental requirement for effective dynamic foveated rendering.

Beyond gaze position alone, cognitive science research indicates that mental workload affects attentional resource allocation and may induce a contraction of the Useful Field of View (UFOV) along with reduced sensitivity to peripheral information [6, 2]. This observation suggests that, in perceptually driven rendering, the optimal spatial extent of the high-quality region should not be treated as a constant but may vary with cognitive state. Under high workload, users may rely less on peripheral information, potentially allowing more aggressive resource allocation strategies; under low workload, a larger high-quality region may be necessary to preserve visual experience [16, 14]. Nevertheless, most existing rendering systems and foveated rendering studies treat gaze position as the primary control variable and rarely incorporate cognitive states such as mental workload into rendering strategies or evaluation protocols [4, 8].

Therefore, moving beyond gaze-only sensing toward the joint perception of gaze and mental workload, and establishing reproducible benchmarks under task-driven, gaze-constrained interaction paradigms, is crucial for exploring more fine-grained and human-centered cognitively adaptive interaction and rendering strategies. The present work follows this direction by providing data, task definitions, and systematic evaluations that lay the groundwork for future methodological and system-level research.

2.4 Summary

Taken together, the three lines of work reviewed above form a tightly coupled system-level loop. On the one hand, mental workload research demonstrates that cognitive state alters attentional allocation and UFOV characteristics, offering an important complementary perspective beyond gaze position alone for perceptually driven optimization. On the other hand, event-based gaze tracking provides a promising sensing foundation for low-latency eye movement perception under head-mounted conditions, yet existing studies largely remain at the level of gaze-only prediction and lack unified evaluation with respect to cognitive state. Meanwhile, the requirements of rendering and interaction systems demand perceptual pipelines that are not only low-latency and robust, but also adaptive to changes in user state, in order to support finer-grained resource allocation and experience optimization.

2.4. Summary

Chapter 3

Experimental Paradigm and Data Collection

This study aims to build a unified perception framework based on an event camera, enabling the simultaneous prediction of gaze coordinates and cognitive load. The goal is to provide reliable inputs for downstream cognitively adaptive resource allocation strategies, such as dynamically adjusting the high-quality region in perception-driven rendering. To support this objective, a high-quality and reproducible dataset is required as the foundation for model training, cross-subject generalization evaluation, and system-level analysis. Therefore, the constructed dataset includes eye event streams, gaze coordinates temporally aligned with the event data, and graded cognitive load labels derived from the N-back task, allowing gaze behavior and cognitive state to be modeled and evaluated within a unified framework.

Specifically, we use the N-back task to manipulate different levels of working-memory load and combine it with a full-screen, target-guided fixation task. Eye event data are recorded under different load conditions. By synchronizing task stimuli and target information, we generate gaze annotations aligned with the event streams and assign cognitive load labels according to task difficulty. The resulting dataset supports two benchmark tasks—gaze prediction and cognitive load classification—and provides a consistent data foundation for subsequent model training, cross-subject generalization analysis, and reliability evaluation.

3.1 Participants

We recruited 15 healthy adult participants (8 female, 7 male). All participants had normal or corrected-to-normal vision and reported no history of neurological or visual disorders. Before the experiment, the participants were informed about the purpose and procedure of the study and gave their written informed consent.

3.2 Overall Experimental Procedure

The experiment was conducted in a fixed screen-based setup. Participants were seated approximately 40 cm away from the display, and all data were collected under identical environmental conditions and hardware configurations. An event camera was rigidly mounted on a desktop stand and aligned with the participant's eye to continuously capture eye-related dynamics (Fig. 3.1).

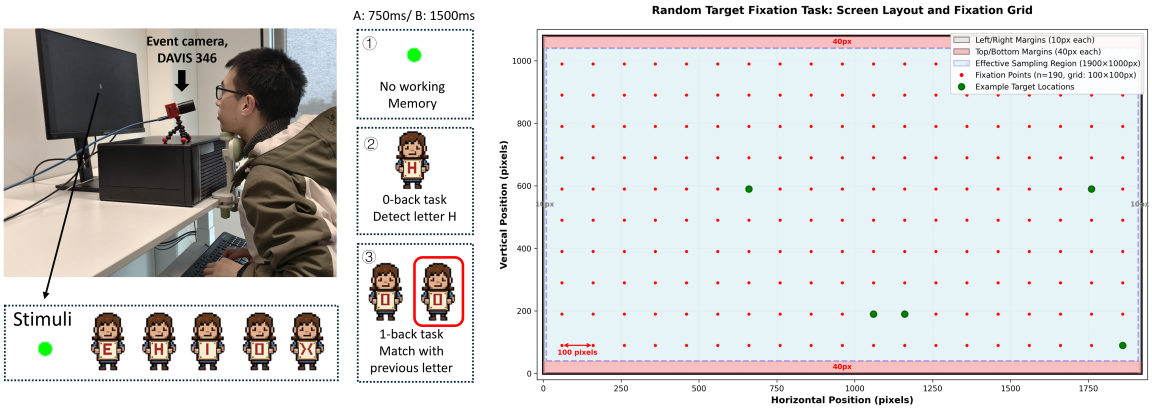

Figure 3.1: Experimental setup and task paradigm for event-based cognitive workload recognition. Left: Recording system with a DAVIS346 event camera capturing high-temporal-resolution eye movements during controlled fixation tasks. Middle: Stimulus design for no-load, 0-back, and 1-back working memory conditions with different presentation inter-stimulus interval (750 ms, 1500 ms). Right: Spatially constrained fixation grid and effective sampling region for the random target fixation task, ensuring consistent gaze behavior across stimuli.

Before the experiment began, participants completed a basic posture calibration procedure. They were instructed to keep their head in a natural position, while a desktop head stabilizer was used to limit large head movements. This setup ensured a stable correspondence between eye behavior and screen coordinates, thereby reducing gaze-estimation errors caused by uncontrolled head motion.

After calibration, participants first performed the fixation task under no-workload

conditions to collect baseline data for gaze-to-screen mapping and to characterize eye behavior in a workload-free state. Subsequently, participants completed fixation tasks under different workload levels, manipulated using an N-back paradigm.

Throughout the experiment, the event camera continuously recorded eye event streams, while task-related behavioral data (e.g., key presses) were logged synchronously. The experiment was divided into multiple short blocks, with mandatory breaks between blocks to mitigate fatigue effects.

3.3 Fixation Task Design

This study employs a full-screen random target fixation task to induce stable and repeatable gaze behavior under controlled conditions. To ensure consistent spatial structure across experimental conditions, we first design a uniform spatial distribution of gaze targets.

Within the screen coordinate system, a set of discrete candidate fixation locations is predefined. To avoid potential display distortion and gaze instability near the screen boundaries, a margin of 10 pixels on the left and right edges and 40 pixels on the top and bottom edges is excluded, resulting in an effective display area of approximately 1900×1000 pixels. Within this region, regular grid sampling is performed at 100×100 pixel intervals, yielding 190 spatially uniform fixation points, as illustrated in the figure 3.2. During the experiment, the target position for each stimulus is randomly selected from this predefined set, ensuring uniform spatial coverage while preventing systematic biases due to positional order. All experimental conditions share this identical spatial configuration.

Based on this spatial design, six experimental conditions are established to systematically manipulate cognitive load and temporal pressure, as summarized in the table 3.1. Each condition consists of 200 stimulus presentations, resulting in 1200 trials per participant in total.

Two conditions serve as no-load baselines. In these conditions, the stimulus is a green dot approximately 15 pixels in diameter (as shown in 3.3(a), used solely to guide gaze movement and fixation without requiring recognition or judgment. The baseline conditions are presented at two stimulus intervals: 1500 ms (slow) and 750 ms (fast), allowing characterization of pure gaze behavior under different levels of time pressure.

In the cognitive-load conditions, the fixation target is replaced by a pixelated stimulus region of approximately 100×100 pixels, containing a letter stimulus of 15×15 pixels. Letters are selected from H, I, O, X, E, as shown in the figure 3.3(b)-(f). These

3.3. Fixation Task Design

Table 3.1: Experimental conditions defined by the combination of cognitive load and presentation speed.

Cognitive load condition	Slow (1500 ms)	Fast (750 ms)
No load	Green fixation dot	Green fixation dot
0-back	Letter judgment (H)	Letter judgment (H)
1-back	Letter judgment (1-back)	Letter judgment (1-back)

letters primarily consist of linear structures and remain highly distinguishable under low-resolution display conditions, thereby reducing potential interference from non-cognitive perceptual factors. Two working-memory paradigms are employed: 0-back and 1-back, each implemented under slow (1500 ms) and fast (750 ms) presentation intervals. In the 0-back condition, participants judge whether the current letter matches the target letter H. In the 1-back condition, participants judge whether the current letter matches the letter presented in the previous stimulus. This design results in four cognitive-load conditions with controlled variations in memory demand and temporal pressure.

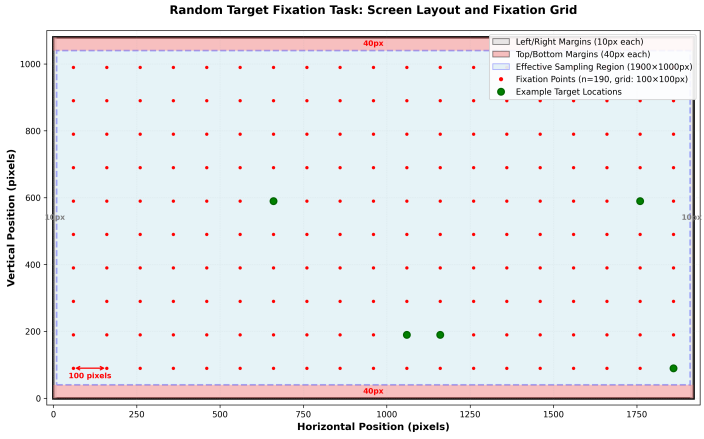

Figure 3.2: Screen layout showing the effective sampling region (1900×1000 px, blue) with excluded margins (gray/red), and the distribution of 190 fixation points (red dots) arranged in a 100×100 px grid. Green dots indicate example randomly selected target locations during experimental stimuluss.

Across all conditions, participants provide responses via key presses: the *Enter* key if the condition is satisfied and the *spacebar* otherwise. The system synchronously records response accuracy and reaction time as behavioral measurements. If a participant's accuracy in a given condition approaches chance level(e.g., below 60% accuracy) , the corresponding data are flagged or excluded in subsequent analyses to ensure data

(a) Fixation target **(b)** E **(c)** H **(d)** I **(e)** O **(f)** X

Figure 3.3: Example visual stimuli used in the N-back task. The fixation target is a green circular dot with a diameter of approximately 15 pixels. Letter stimuli are presented as a cartoon character holding a single uppercase letter (E, H, I, O, X), where the letter glyph size is approximately 15 pixels.

validity.

Through this design, the six experimental conditions form a discrete progression from no load to higher cognitive load, serving as categorical labels for subsequent model training and evaluation. Further details of the dataset are summarized in Table 3.2.

Table 3.2: Summary statistics of the EveLoad dataset.

Item	Description
#Subjects	15 (8 female, 7 male)
Recording time per subj.	22.5 min
Total recording time	5.6 h
#Stimuli	18,000
#Level (#Class)	6 (No-load/0-back/1-back × Slow/Fast)
Annotation granularity	Stimulus-level
Labels per stimulus	Stimulus coordinate, workload level, response

3.4 Data Acquisition and Synchronization

During the experiment, the event camera asynchronously recorded pixel-level brightness changes in the eye region, producing a high-temporal-resolution event stream. In parallel, task-related information—including fixation target positions, stimulus presentation timing, and behavioral responses—was recorded. All data streams were synchronized using a unified timestamp reference shared with the event camera, ensuring precise temporal alignment between eye events and task states.

As a result, each stimulus corresponds to a continuous segment of event-based eye movement data associated with a specific fixation target and mental workload label, forming the basis for subsequent gaze estimation and mental workload recognition

experiments.

3.5 Stimulus-Level Annotation

Each fixation task instance was defined as an independent stimulus and annotated using a structured label file. Annotations were organized at the stimulus level and described fixation target properties, mental workload conditions, and participant behavior. Each record includes the fixation target's screen coordinates (x, y), the presented letter stimulus and whether it was a target (`letter`, `is_target`), the participant's response and correctness (`user_response`, `is_correct`), and the stimulus start and end timestamps (`start_time_us`, `end_time_us`). In addition, stimulus duration statistics (`duration_us`, `duration_ms`, `stimulus_event_count`) and participant and condition identifiers (`participant_id`, `condition`, `block`, `stimulus`) were recorded.

Based on these annotations, each stimulus can be precisely aligned with a corresponding segment of event-based eye movement data and assigned a unique mental workload label (e.g., 0-back or 1-back). This stimulus-level annotation scheme provides a unified and reproducible interface for subsequent gaze estimation and mental workload recognition analyses.

Chapter 4

Data Quality Assessment and Dataset Curation

This chapter provides a systematic quality assessment and curation of the constructed event-based eye-tracking dataset, which serves as the foundation for all subsequent modeling and analysis. Rather than optimizing model performance, the focus here is to validate the dataset from an experimental and methodological perspective, ensuring temporal continuity, correct label alignment, and the exclusion of samples that violate experimental assumptions. In addition, the chapter examines whether low-level statistical properties inherent to event-based sensing—particularly event rate—exhibit systematic variation across experimental conditions and could therefore act as confounding cues. Accordingly, this chapter is organized into two stages: (1) Data Quality Assessment, which identifies and removes invalid or anomalous samples, and (2) Confounding Cue Analysis, which evaluates the potential influence of condition-correlated sensory statistics. Together, these analyses establish a validated and interpretable data foundation for subsequent model training and cognitive load evaluation.

4.1 Data Quality Assessment

This section presents a systematic quality assessment of the event-based eye-tracking dataset, focusing on acquisition reliability and supervision validity. Before conducting task-level analyses or model training, we verify that the event streams are temporally continuous, properly aligned with supervisory signals, and free from structural

anomalies that could bias supervised learning.

Unlike frame-based eye-tracking data, event cameras asynchronously record brightness changes as continuous event streams without explicit frame boundaries. While this representation provides high temporal resolution, it also increases sensitivity to timestamp irregularities, recording interruptions, and misalignment between events and labels. Such issues may not be evident from global statistics but can still affect model behavior.

Accordingly, we evaluate data quality from three complementary perspectives: (1) temporal continuity of the event stream, (2) alignment between event data and gaze supervision, and (3) detection and handling of spatially abnormal samples. These steps ensure that subsequent analyses are conducted on structurally consistent and interpretable data.

4.1.1 Temporal Continuity

The first step of data quality assessment is to examine the temporal integrity of the event stream. Under normal acquisition conditions, an event camera should continuously generate events throughout task execution. Extended periods without events typically indicate sensor interruptions, data transmission failures, or recording thread stalls, all of which directly compromise temporal continuity.

To assess temporal continuity, we analyzed the time intervals between consecutive events within each recording, with particular focus on the maximum inter-event gap (*max gap*) and the occurrence of abnormal temporal dropouts. Specifically, the maximum gap between adjacent events was computed at the stimulus level, and worst-case values and percentile statistics were aggregated at the condition level to identify potential structural gaps in the temporal domain. stimulus exhibiting gaps exceeding a predefined threshold of 100 ms were further inspected to trace their origin and handled in subsequent bad-sample processing. As shown in Fig. 4.1(left), the maximum inter-event interval across recordings remains at a generally low level under all experimental conditions. Only a very small number of recordings exhibit abnormal temporal gaps exceeding 100 ms, and these outliers do not show systematic clustering with respect to task load level or stimulus presentation speed. This result indicates that, even in the most extreme cases, the event streams do not contain long temporal interruptions that would violate the experimental assumptions, and temporal continuity is largely preserved at the recording level.

It should be noted that the maximum inter-event interval only reflects the presence

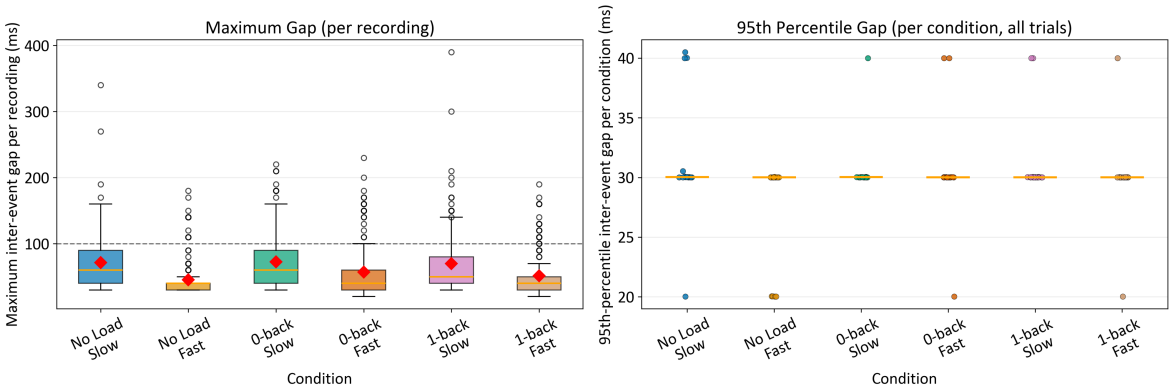

Figure 4.1: Temporal continuity assessment. (Left) Maximum gap, (Right) 95th-percentile gap.

of extreme temporal gaps. Even when the maximum interval of all recordings remains below an abnormal threshold, the overall temporal rhythm of the event stream may still undergo systematic changes at the distribution level, for example, a global increase in inter-event intervals without reaching the level of a true interruption. To exclude the possibility of a non-interrupted but low-density event stream, we further analyze the 95th percentile of the inter-event interval distribution.

As shown in Fig. 4.1 (right), the 95th percentile inter-event intervals across recordings are highly consistent across experimental conditions, remaining stably concentrated around approximately 30 ms and exhibiting no systematic drift with task load or stimulus speed. This result indicates that, after excluding extreme outliers, the event generation rhythm is largely preserved across conditions for the vast majority of time.

It is important to note that the 95th percentile interval characterizes the upper tail of the inter-event interval distribution rather than the mean event rate. In stable fixation tasks, event camera outputs typically exhibit a burst-like temporal structure, where dense events are generated during eye movements or micro-movements, while event activity becomes relatively sparse during fixation maintenance. As a result, even when nearly every 10 ms window contains valid events, high-percentile inter-event intervals may still reach tens of milliseconds. The stability of the 95th percentile across conditions indicates that this burst-like temporal structure remains consistent and does not undergo systematic changes between experimental conditions.

Overall, the event data exhibit strong temporal continuity, and the vast majority of recordings do not contain interruptions that would compromise analytical validity. Although isolated abnormal gaps are observed in a small number of recordings,

their proportion is limited and no systematic dependence on subject or experimental condition is observed. Given that a certain degree of sensing and acquisition noise is unavoidable in practical event-based recordings, these isolated anomalies are retained in the dataset and their potential impact is absorbed through statistical aggregation and learning-based modeling rather than enforced removal.

4.1.2 Alignment Validity

After confirming overall temporal continuity, it is necessary to verify the validity of temporal alignment between the event data and supervision labels at the stimulus level. In this study, all data were generated using a unified acquisition and annotation pipeline, and both fixation targets and experimental condition labels remain constant throughout each stimulus. Nevertheless, explicit verification on the exported dataset is required to ensure that event–label correspondence satisfies the basic assumptions of supervised learning.

Accordingly, three minimal yet sufficient alignment checks were conducted at the stimulus level. First, we verified that each stimulus contains complete supervision metadata, including subject identity, experimental condition, stimulus index, and valid start and end timestamps. Second, we examined the number of events per stimulus to confirm that each stimulus presentation interval contains valid event input, and to identify any stimulus with abnormally low event counts. Third, temporal boundary consistency was validated by ensuring that recorded durations match timestamp differences, thereby excluding segmentation errors.

Across all **20,400** stimulus, no missing supervision fields, inconsistent temporal boundaries, absent events, or low-density stimulus were detected. These results indicate that the event data maintain consistent temporal-semantic correspondence with static fixation targets at the stimulus level, and no systematic event–label misalignment is present.

4.1.3 Bad Sample Detection and Removal

Beyond temporal continuity and label alignment, event-based data may also be affected by sensor-level spatial anomalies, most notably high-frequency events persistently triggered by a small number of pixels (hot pixels). Such anomalies are highly localized in space and unrelated to genuine eye-movement behavior. If present in sufficient quantity, they may interfere with event-based statistical analysis or model training, and therefore require basic inspection.

In this study, hot-pixel detection is primarily conducted through visualization of the spatial event distribution. Specifically, an accumulated event heatmap is generated for each recording, and the spatial distribution is inspected individually to identify abnormal event concentrations dominated by single pixels. Through this procedure, only a very small number of hot pixels were observed, with a total count of approximately six across all recordings. Moreover, these hot pixels did not exhibit any systematic clustering with respect to specific subjects or experimental conditions, indicating the absence of sensor-level systematic spatial anomalies.

For the identified hot pixels, a simple spatial masking strategy is applied to remove events generated at the corresponding pixel locations. This operation acts solely on the spatial dimension and does not alter the temporal structure or polarity information of the event stream. A comparison of event-rate distributions before and after hot-pixel masking across different experimental conditions shows that the overall distribution shape and relative relationships between conditions remain unchanged, with no observable systematic shift introduced by the masking procedure. Based on these results, all subsequent analyses and modeling are conducted on the hot-pixel-masked data, while the original data are retained for robustness verification.

4.2 Confounding Cue Analysis

After confirming structural validity in terms of temporal continuity, label alignment, and spatial integrity, the dataset satisfies the basic experimental assumptions. However, structural correctness does not preclude the presence of potential confounding factors during model learning. In event-based sensing, models may implicitly exploit low-level statistical properties that correlate with experimental conditions, even in the absence of explicit acquisition errors.

In the current design, combinations of memory load and stimulus presentation speed may influence global event statistics, such as event rate. Although such differences are not anomalies, systematic correlations with task conditions could bias model interpretation. Therefore, potential confounding cues must be examined prior to task-specific modeling.

To this end, we analyze event rate as a representative low-level statistic. We first compare event-rate distributions across conditions at the cross-subject level, and then conduct subject-wise normalized analyses to assess within-subject consistency and stability. This evaluation determines whether event rate forms a stable condition-related structure in the dataset.

4.2.1 Event Rate Distribution Across Conditions

After completing data quality control and removing non-behavioral spatial artifacts such as hot pixels, we examine the statistical behavior of event rate across experimental conditions. Event rate is defined as the number of events generated per unit time (events/s), reflecting the overall responsiveness of the event camera to retinal motion and local luminance changes. As a spatially agnostic low-level statistic, event rate can be readily exploited by learning models, making it necessary to first characterize its distributional properties under different memory load and stimulus speed conditions.

Figure 4.2 shows the cross-subject distributions of event rate under the six experimental conditions (No Load, 0-back, 1-back × Slow / Fast). Substantial overlap can be observed across all conditions, with no clear separable boundaries in absolute value. Within-condition variability is large, and the magnitude of cross-subject variation exceeds the differences between condition-wise means or medians, indicating that individual differences dominate the absolute scale of event rate.

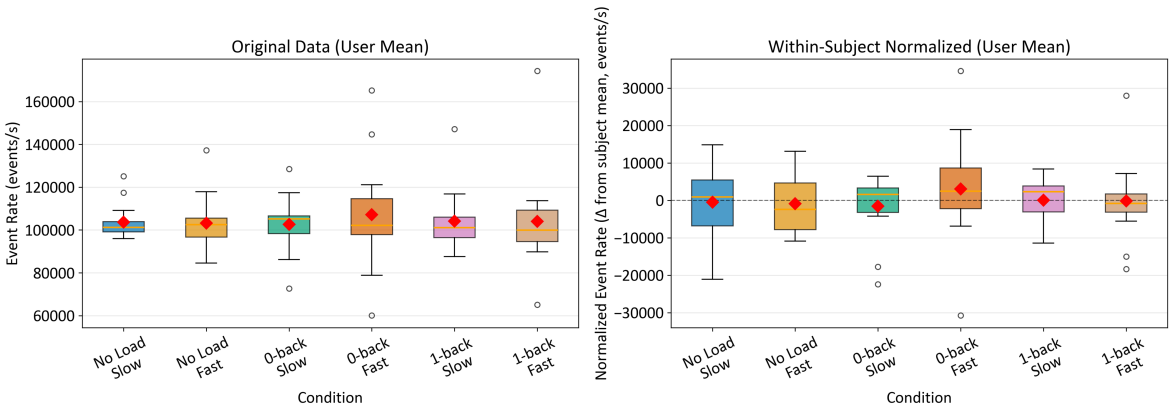

Figure 4.2: Effect of within-subject normalization on event rate distributions across experimental conditions. Left: raw event rates. Right: within-subject normalized event rates.

From the relative positions of condition-wise central tendencies, no consistent monotonic trend across memory load levels is observed. For instance, the 0-back condition exhibits slightly higher event rates under some speed settings, while the 1-back condition tends to occupy an intermediate range. The No Load condition occasionally corresponds to lower event rates, but this pattern is not consistently present across subjects. Similarly, differences between Slow and Fast conditions are modest and do not scale proportionally with stimulus presentation speed.

It is important to note that event rate reflects event density per unit time rather than the total number of events per stimulus. Since Slow and Fast conditions differ

substantially in stimulus duration, total event counts are not directly comparable across these conditions. Event rate therefore serves as a normalized measure that avoids confounding effects introduced by differences in temporal window length.

Overall, event rate does not form a clearly separable or generalizable structure across experimental conditions at the cross-subject level. Although small systematic shifts can be observed in condition-wise central tendencies, these effects are largely masked by inter-subject variability. Cross-subject distributions alone are therefore insufficient to determine whether event rate encodes stable condition-related information, motivating further analysis at the within-subject level.

4.2.2 Information Stability Analysis

The preceding analysis indicates that absolute event-rate values are strongly influenced by individual differences. However, this observation does not exclude the possibility that event rate may respond to condition changes in a consistent manner within individual subjects. To assess this possibility, event rate is further analyzed at the within-subject level.

For each subject, event rates under different experimental conditions are normalized by subtracting the subject-specific mean. This transformation removes differences in overall event generation levels between subjects and focuses the analysis on relative changes induced by experimental conditions.

Figure 4.2 shows the distributions of within-subject normalized event rates across all conditions. For all conditions, the distributions are centered around zero, with both medians and means remaining close to zero. No condition exhibits a systematic positive or negative shift relative to the subject-specific baseline, indicating the absence of a consistent directional response.

Comparisons across conditions reveal similar dispersion and value ranges. No condition consistently shows higher or lower normalized event rates within subjects. Although some subjects exhibit pronounced deviations under specific conditions, the directions of these deviations are not consistent across subjects and do not form a shared pattern.

These results indicate that, after removing subject-level baseline differences, event rate lacks stable and consistent responses to experimental condition changes at the within-subject level. Condition-related effects in event rate primarily manifest as subject-specific fluctuations rather than a unified condition-driven structure. Without introducing explicit discriminative modeling, event rate alone does not support reliable

differentiation of experimental conditions. The implications of this finding are further discussed in subsequent sections in conjunction with modeling results.

Chapter 5

Methods

This chapter presents the methodology for two learning tasks constructed from asynchronous eye-tracking event streams: (i) gaze estimation, formulated as continuous coordinate regression, and (ii) cognitive load prediction, formulated as six-class classification. Although both tasks share the same raw event source, their supervision structure and modeling objectives differ, resulting in task-specific input representations and model designs.

We describe the preprocessing pipeline, model architectures, loss functions, and experimental protocols for both tasks, forming the methodological basis for subsequent analysis.

5.1 Data Preprocessing

This section describes the preprocessing pipeline that transforms raw event streams into trainable samples. Since gaze estimation and cognitive load prediction use different supervision signals and objectives, their sample construction and data partitioning strategies differ. We therefore introduce the task-specific label processing, event representation, and split protocols used in this work.

5.1.1 Data Preprocessing for Gaze Estimation

This section introduces the data processing workflow for the gaze point estimation task. Because this task has high requirements for temporal continuity and spatial positioning accuracy, data preprocessing involves not only the construction and filtering of the

original event stream, but also the temporal alignment of supervision labels and sample generation strategies. We describe gaze label interpolation, event window construction and filtering, and the within-subject split protocol used for training and evaluation.

1. Gaze Label Processing

In the gaze estimation task, each stimulus corresponds to a discrete target coordinate (x, y) on the display and is associated with a start timestamp t_s and an end timestamp t_e. Since target locations change abruptly across stimuli, directly assigning a constant label throughout each stimulus would introduce temporal discontinuities in the supervision signal at stimulus boundaries. Such discontinuities are inconsistent with the continuous nature of eye movement, as subjects typically undergo a transition phase when shifting gaze from the previous fixation point to the new target.

To alleviate this discontinuity introduced by stimulus-level discrete labels, a temporally continuous supervision signal is constructed by applying linear interpolation at the beginning of each stimulus. Specifically, during the initial $T_{\text{interp}} = 100\,\text{ms}$ interval following stimulus onset, the supervision signal gradually transitions from the previous stimulus target location to the current target location in a linear manner. After this interpolation interval, the label remains constant at the current stimulus coordinate for the remainder of the stimulus.

For the first stimulus in a sequence, where no preceding target exists, the label is directly set to the current stimulus target without interpolation.

During sample generation, each event window is assigned a supervision label by aligning the window end time with the interpolated gaze signal along the temporal axis. This strategy ensures temporal consistency while preserving causality, as the supervision assigned to each window depends only on information available up to the prediction time.

To assess the influence of interpolation on model performance, an additional training setting without label interpolation is included as a control condition for comparison.

2. Event Data Processing

This section describes the event processing pipeline adopted for the gaze estimation task. Since accurate gaze prediction requires both high spatial precision and strict temporal consistency, the preprocessing procedure is designed to suppress noise while preserving event structures that are informative for eye movement dynamics. Given the highly transient nature of saccades and fixation transitions, the temporal structure of

the event stream must be maintained under a strict causal constraint. Accordingly, the processing pipeline consists of stimulus-level valid sampling, causal temporal window construction with weighted filtering, and event frame representation.

stimulus-Level Valid Sampling Region To avoid cross-stimulus contamination and low-information event segments, samples are generated only from a fixed prefix interval following stimulus onset. The valid sampling duration is set to $T_{\text{valid}} = 500\,\text{ms}$, and event data outside this interval are excluded from gaze sample construction. This constraint reduces the influence of late-stage fixation periods where event activity may be less informative for gaze localization.

Causal Temporal Windowing and Weighted Filtering To suppress windows dominated by peripheral noise while emphasizing recent eye dynamics, we compute a weighted event contribution score using temporal decay and spatial center weighting. Windows whose score is below a threshold $S_{\min}$ are discarded before representation construction. The same decay constant is used for both window scoring and event accumulation to keep the filtering criterion consistent with the final representation.

Event Frame Representation For retained windows, events are accumulated into a two-channel, polarity-separated event frame representation $\mathbf{X} \in \mathbb{R}^{2\times H\times W}$. Temporal decay is applied during accumulation to emphasize recent events. The accumulated values are clipped at an upper bound L and normalized to a fixed numerical range to ensure consistent distribution across windows and improve training stability.

The key hyperparameters used in event windowing and frame construction are summarized in Table 5.1.

Table 5.1: Key parameters used in event windowing and event frame construction.

Parameter	Value	Description
Window length T_w	$10\,\text{ms}$	Length of causal temporal window
Step size Δ	$5\,\text{ms}$	Sliding window step size
Temporal decay constant τ	$5000\,\mu\text{s}$	Temporal decay weighting parameter
Spatial scale σ	$0.2 \cdot \min(H, W)$	Gaussian spatial center weighting
Weighted threshold $S_{\min}$	200	Weighted event contribution threshold
Clipping limit L	15	Upper bound for event frame values
Resolution $H \times W$	260×346	Event camera spatial resolution

3. Split Strategy

In the within-subject setting, a separate model is trained per participant using only that subject's samples. To prevent temporal leakage, stimuli are partitioned as the minimum split unit rather than individual windows. To assess generalization within the same individual while preventing temporal data leakage, samples are partitioned at the stimulus level rather than at the window level.

Specifically, stimuli are randomly divided into training, validation, and test sets with a ratio of 8:1:1. All event windows derived from the same stimulus are assigned to the same subset, ensuring that temporally adjacent or overlapping samples do not appear across different splits.

This strategy preserves the temporal integrity of each stimulus and provides a consistent evaluation protocol for subject-specific gaze modeling.

5.1.2 Event Data Processing for Cognitive Workload Prediction

This section describes the event data processing procedure adopted for the cognitive workload prediction task. Unlike gaze estimation, which requires fine-grained temporal alignment with continuous supervision, workload prediction focuses on discriminative event statistics and short-term temporal dynamics. Accordingly, event representation is constructed using fixed event-count aggregation rather than fixed-duration time windows.

Event-Count-Based Frame Construction The raw event stream is segmented according to a fixed number of events. Specifically, every N_e consecutive events are accumulated to form a two-channel, polarity-separated event frame. This strategy ensures that each frame contains a comparable amount of information, thereby reducing variability caused by fluctuations in event rate.

To encode short-term temporal dynamics, T consecutive event frames are stacked along the channel dimension to form the final input tensor

$$\mathbf{X} \in \mathbb{R}^{(2T)\times H\times W}. \tag{5.1}$$

In this work, we set $N_e = 1000$ events per frame and $T = 10$ frames per sample, resulting in $2T = 20$ input channels. Temporal decay, value clipping, and normalization follow the same configuration as described in the gaze estimation preprocessing

stage.

Evaluation Protocol and Data Splitting All samples are generated offline during preprocessing. To evaluate model performance under different generalization settings, three splitting protocols are considered.

Within-subject setting. For each subject, stimuli are treated as the minimum partitioning unit and are divided into training, validation, and test sets with a fixed ratio of $8 : 1 : 1$. All samples derived from the same stimulus are assigned to the same subset, thereby preventing temporal leakage caused by overlapping or adjacent frames.

Cross-subject setting. Subject identity is used as the minimum split unit. In each split, one subject is reserved for validation, another distinct subject is reserved for testing, and the remaining subjects jointly form the training set. All stimuli belonging to a subject are assigned to a single subset, ensuring that no subject appears simultaneously in training and evaluation stages.

Pooled multi-subject setting. For each subject, stimuli are first partitioned into training, validation, and test subsets using the same $8 : 1 : 1$ ratio. After subject-wise partitioning, all training subsets are merged to form a unified training set, and the same merging procedure is applied to validation and test subsets. Under this protocol, the model is trained on data from multiple subjects simultaneously while preserving stimulus-level integrity within each individual.

5.2 Model Design Space

Due to the different output spaces and supervision signals, we adopt task-specific model designs. Gaze estimation predicts continuous coordinates in display space, while workload prediction outputs six-class logits. This section describes the gaze estimation model and the workload classification model used in experiments.

5.2.1 Gaze Estimation Model

The overall architecture of the gaze estimation model is illustrated in Fig. 5.1. The network follows an encoder–neck–prediction head structure, where event-based inputs are progressively transformed into spatial response maps and an associated uncertainty estimate.

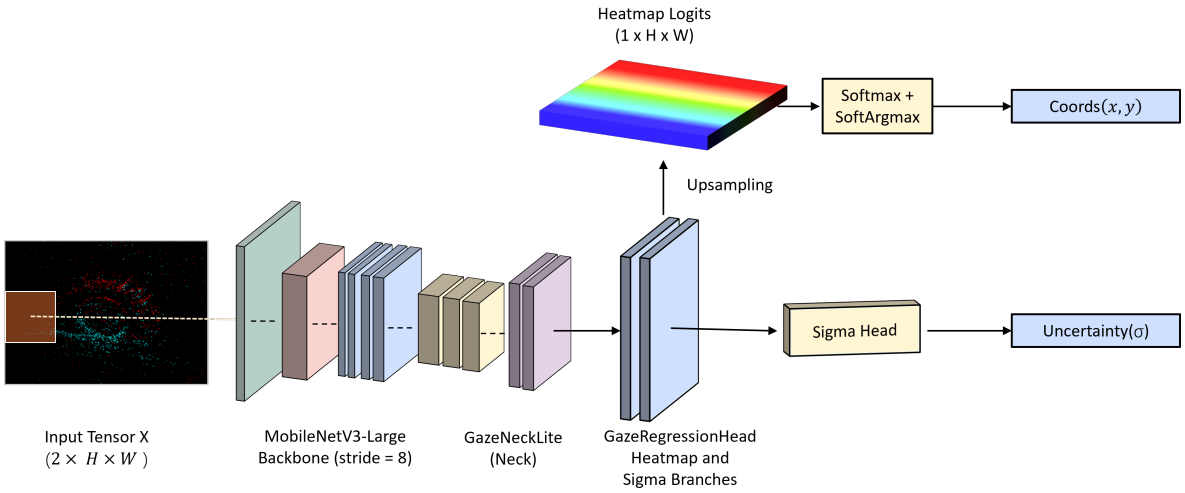

Figure 5.1: Gaze estimation model architecture. The network processes a two-channel event frame using a MobileNetV3-Large backbone and GazeNeckLite module. Parallel heatmap and sigma branches predict gaze coordinates and uncertainty, respectively.

Input. As shown in Fig. 5.1, each gaze sample consists of a single two-channel event frame constructed from a causal temporal window:

$$\mathbf{X} \in \mathbb{R}^{B\times 2\times H\times W}. \tag{5.2}$$

The two channels correspond to positive and negative event polarities.

Backbone. The input tensor is first processed by a MobileNetV3-Large backbone, which extracts hierarchical spatial features from the event representation. The first convolutional layer is adapted to accept two-channel input, and the output stride is set to 8 to preserve sufficient spatial resolution for precise localization.

Neck. The backbone features are then fed into a lightweight feature fusion module, *GazeNeckLite*, as depicted in Fig. 5.1. This module aggregates multi-scale information into a high-resolution representation. Coordinate encoding and lightweight channel attention mechanisms are incorporated to enhance spatial awareness and improve localization accuracy.

Prediction Head and Decoding. Finally, the fused feature map is passed to the regression head, which contains two parallel branches. The heatmap branch produces a single-channel spatial response map (heatmap logits), which is upsampled to the original spatial resolution. The logits are converted into a spatial probability distribu-

tion via softmax, and continuous gaze coordinates are obtained through soft-argmax decoding.

In parallel, a sigma branch predicts an uncertainty scale parameter σ, providing an explicit estimate of prediction reliability, as illustrated in Fig. 5.1.

5.2.2 Cognitive Workload Classification Model

The overall architecture of the cognitive workload classification model is illustrated in Fig. 5.2. The network processes stacked event-frame representations and produces six-class logits corresponding to workload conditions.

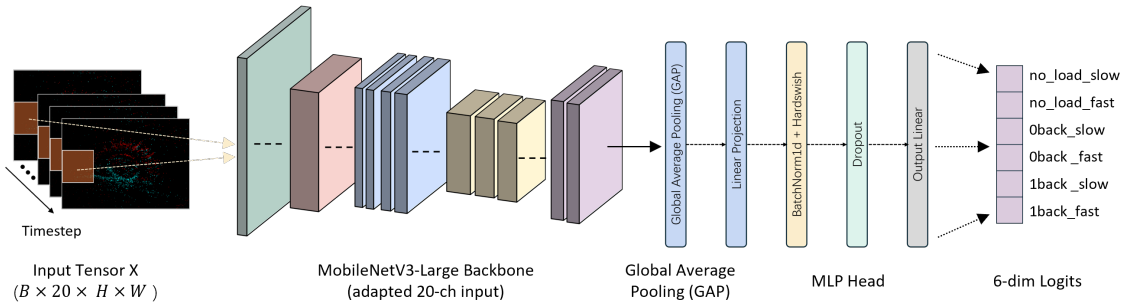

Figure 5.2: Architecture of the cognitive workload classification model. Ten consecutive two-channel event frames are stacked to form a 20-channel input tensor. The adapted MobileNetV3-Large backbone extracts spatial features, followed by Global Average Pooling (GAP) and an MLP classification head that outputs six-class logits corresponding to workload conditions.

Input. As shown in Fig. 5.2, each workload sample consists of a 20-channel tensor formed by stacking 10 consecutive two-channel event frames:

$$\mathbf{X} \in \mathbb{R}^{B\times 20\times H\times W}. \tag{5.3}$$

Backbone. The input tensor is processed by a MobileNetV3-Large backbone, with the first convolutional layer adapted to accept 20-channel input. To leverage ImageNet-pretrained weights, the original 3-channel convolution kernels are averaged across channels, replicated to match the 20-channel dimensionality, and scaled to preserve output magnitude.

Classification Head. The extracted backbone features are first reduced by Global Average Pooling (GAP), producing a compact feature vector. This vector is then passed to an MLP head consisting of a linear projection, normalization layer, Hardswish

activation, Dropout, and a final linear layer that outputs six-dimensional logits corresponding to the workload categories.

5.3 Loss Functions

This section describes the optimization objectives used for the two tasks. Since gaze estimation and cognitive workload prediction differ in output space and supervision form, task-specific loss functions are adopted. Gaze estimation is formulated as a continuous regression problem with uncertainty modeling and spatial supervision, whereas workload prediction is treated as a multi-class classification problem optimized using cross-entropy loss.

5.3.1 Gaze Estimation Loss

The gaze estimation objective jointly constrains coordinate accuracy and spatial response consistency. The overall loss function is defined as

$$\mathcal{L} = w_{xy}\,\mathcal{L}_{xy} + w_{\text{heat}}\,\mathcal{L}_{\text{heat}}, \tag{5.4}$$

where $\mathcal{L}_{xy}$ denotes the coordinate regression loss and $\mathcal{L}_{\text{heat}}$ denotes the heatmap supervision loss.

Coordinate Regression Loss. The coordinate loss combines Smooth L1 regression with a Gaussian negative log-likelihood (GNLL) term:

$$\mathcal{L}_{xy} = (1 - w)\,\mathcal{L}_{\text{smooth}} + w\,\mathcal{L}_{\text{gnll}}. \tag{5.5}$$

The Smooth L1 term penalizes the deviation between the predicted coordinate $\hat{\mathbf{x}}$ and the ground-truth coordinate $\mathbf{x}$:

$$\mathcal{L}_{\text{smooth}} = \text{SmoothL1}(\hat{\mathbf{x}}, \mathbf{x}). \tag{5.6}$$

To incorporate uncertainty estimation, the GNLL term is defined as

$$\mathcal{L}_{\text{gnll}} = \frac{d^2}{2\sigma^2} + 2\log(\sigma), \tag{5.7}$$

where d denotes the Euclidean distance between $\hat{\mathbf{x}}$ and $\mathbf{x}$, and σ is the predicted uncertainty scale.

Heatmap Supervision Loss. To encourage spatially coherent response distributions, a target heatmap H is constructed as a two-dimensional Gaussian centered at the ground-truth location. The predicted heatmap $\hat{H}$ is supervised using focal loss:

$$\mathcal{L}_{\text{heat}} = \text{FocalLoss}(\hat{H}, H). \tag{5.8}$$

5.3.2 Workload Classification Loss

The cognitive workload prediction task is formulated as a six-class classification problem. Given model logits $\mathbf{z}_i \in \mathbb{R}^6$ and ground-truth label $y_i \in \{0, \ldots, 5\}$, the optimization objective is the standard cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(z_{i,y_i})}{\sum_{c=1}^{6} \exp(z_{i,c})}. \tag{5.9}$$

5.4 Experimental Configurations

This section specifies the evaluation protocols and training settings for gaze estimation and cognitive workload prediction. We report the data partition strategy, input construction, model configuration, and optimization details. Gaze estimation is evaluated under a within-subject protocol, whereas workload prediction is evaluated under within-subject, cross-subject, and pooled multi-subject settings.

5.4.1 Gaze Estimation Experiments

Gaze estimation experiments follow a within-subject protocol, where one model is trained per participant. Stimuli are partitioned into training, validation, and test sets with a fixed ratio of 8:1:1.

Input and Window Configuration. Event frames are constructed from causal temporal windows of 20 ms with a step of 10 ms. Only the first 500 ms after stimulus onset is used for sample generation. Windows with fewer than 200 events are discarded. Temporal decay uses $\tau = 5000\,\mu$s, values are clipped at $L = 15$ before normalization, and spatial center weighting is enabled (Gaussian scale factor 0.25).

Model and Optimization. The model follows Section 5.2.1 (MobileNetV3-Large, stride 8; GazeNeckLite with 160 channels and two blocks; head mid-channel 96; soft-argmax temperature 0.35). Training uses Adam for 300 epochs with learning rate

1.5×10^{-3}, weight decay 1×10^{-4}, and batch size 32. Gradient clipping (max norm 0.5) and AMP are enabled. All runs use seed 42.

Loss. Coordinate regression and heatmap supervision are equally weighted ($w_{xy} = 1.0$, $w_{\text{heat}} = 1.0$). The heatmap loss uses MSE. GNLL is implemented but disabled in these experiments (weight set to zero).

5.4.2 Cognitive Workload Prediction Experiments

Workload prediction uses the event-count-based representation in Section 5.1.2. Each sample stacks 10 consecutive frames built from 1000 events per frame (20 channels). Temporal decay uses $\tau = 10{,}000\,\mu$s, and values are clipped by a percentile-based threshold before normalization.

Evaluation Protocols. Three settings are considered. (i) *Within-subject:* one model per participant; stimuli are split into train/val/test with 8:1:1. (ii) *Cross-subject:* subjects are the minimum split unit; one subject is used for validation, a different subject for testing, and the remaining subjects for training. (iii) *Pooled multi-subject:* stimuli are first split per subject with 8:1:1, then merged across subjects within each split to form pooled train/val/test sets.

Model and Optimization. The model follows Section 5.2.2 (MobileNetV3-Large backbone, stride 8, adapted to 20-channel input; GAP + MLP head). Training uses Adam for up to 20 epochs with learning rate 1×10^{-4}, weight decay 1×10^{-4}, and batch size 4. Gradient clipping (max norm 0.5) is applied and seed 42 is used. Early stopping monitors validation loss with patience 3. AMP is disabled.

Chapter 6

Experiment results

This chapter presents the empirical evaluation of the proposed methods on both gaze estimation and cognitive load classification tasks. We first report quantitative performance under different supervision and data-split settings, and then provide detailed analyses to examine stability, generalization behavior, and underlying statistical characteristics of the event data.

6.1 Gaze Estimation Performance

Under the current dataset and supervision setting, the single-frame event-based CNN failed to achieve stable and generalizable gaze regression performance.

On the validation set, the average L2 error remained around 450 pixels, which is substantially higher than the level required for precise screen-space gaze estimation. Across different training runs, validation performance exhibited noticeable fluctuations, indicating instability in convergence.

Although the training loss consistently decreased over epochs, the validation error did not show corresponding improvement. This divergence suggests that the model was unable to learn a mapping that generalizes beyond the training samples.

To verify the correctness of the training pipeline, we conducted overfitting experiments on highly reduced subsets of the data. In these controlled settings, the model was able to fit the training samples effectively, confirming that the network architecture, loss formulation, and coordinate transformation were implemented correctly. Therefore, the poor validation performance cannot be attributed to implementation errors.

We further introduced label interpolation to construct continuous gaze trajectories between discrete target points. However, this modification did not lead to measurable improvement in validation accuracy.

Overall, under the present formulation—characterized by single-frame event inputs and screen-coordinate regression supervision—the model did not exhibit stable or satisfactory generalization performance.

6.2 Cognitive Load Classification Performance

This section evaluates cognitive load classification under three experimental protocols: within-subject training, cross-subject generalization, and pooled multi-subject training. These settings allow us to assess both subject-specific discriminative capability and cross-individual transferability of the learned representations.

6.2.1 Within-Subject Performance

Figure 6.1 presents the distribution of classification accuracy on the test set under the within-subject setting. It can be observed that the model achieves relatively high classification accuracy for most participants, with overall performance concentrated in a high accuracy range. Except for a small number of subjects, the test accuracy for most users approaches or exceeds 95%, indicating that under within-subject conditions, the model can reliably distinguish between different cognitive workload levels.

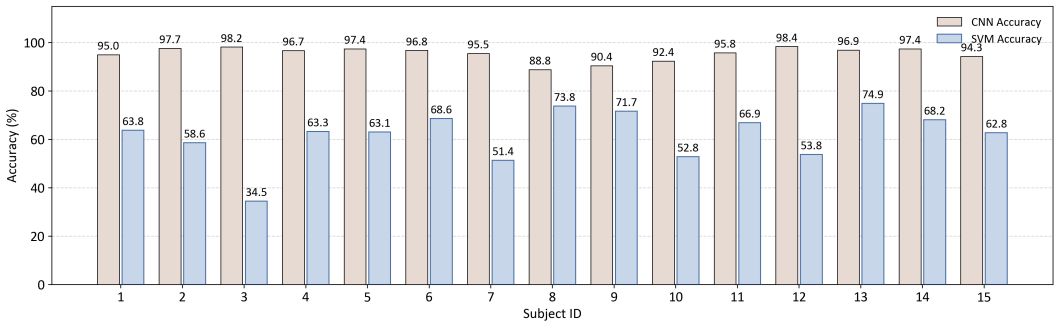

Figure 6.1: Subject-specific classification accuracy on the test set for 15 participants. The accuracy of the proposed CNN model (95.5% on average) is compared with a linear SVM based on low-level event statistics (61.2% on average).

6.2.2 Cross-Subject Generalization

Under the cross-subject setting, training and testing data were strictly separated at the participant level. The overall classification accuracy reaches approximately 39%, which is above the chance level of a six-class task (16.7%) but remains substantially lower than within-subject performance. Accuracy varies across test participants but stays within a relatively limited range. The pronounced performance drop indicates limited cross-individual generalization under the current input representation and supervision design.

6.2.3 Pooled Multi-Subject Training

Under the pooled multi-subject setting, data from all participants were combined and then partitioned into training and testing sets at the stimulus level. In this configuration, samples from the same participants may appear in both training and testing sets, although no individual stimulus overlaps across splits.

In contrast to the cross-subject setting, the model achieves substantially higher performance under pooled training. The overall classification accuracy reaches approximately 90%, indicating that the network is able to learn discriminative patterns when trained on a mixed distribution of subjects.

Compared with the within-subject results, performance remains competitive, and no significant instability is observed during training. These findings suggest that, under shared subject distributions between training and testing data, the model can effectively capture task-relevant features.

6.3 Result Analysis

This section analyzes the observed performance patterns from both tasks. We examine potential causes of instability in gaze regression and investigate the statistical characteristics underlying cognitive load classification results, with particular attention to subject-specific effects and generalization limitations.

6.3.1 Gaze Estimation Analysis

Under the current dataset configuration, the gaze prediction task exhibits pronounced instability. When evaluated on the full dataset, the validation error consistently remains at a high level, and no stable or sustained downward trend is observed during

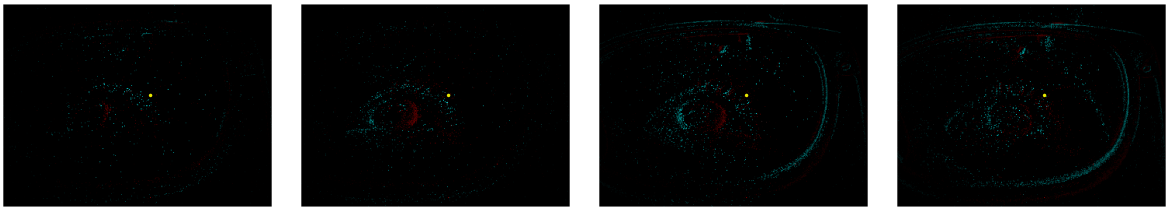

Figure 6.2: Event frames corresponding to the same gaze coordinate label. The yellow point marks the annotated gaze position. Notice that substantial variations in eye orientation are observed despite identical coordinate annotations.

training. Despite extended optimization and repeated tuning of hyperparameters, the validation error continues to fluctuate within a relatively large range and eventually saturates without meaningful improvement. This behavior indicates limited generalization capability under the present setting.

To rule out the possibility that the observed performance limitation is caused by insufficient model capacity or suboptimal training strategies, we systematically evaluated multiple model variants and optimization configurations, including different backbone architectures, loss formulations, and temporal input constructions. Across these configurations, no significant or consistent improvement in validation performance was observed. Although certain settings reduced the training error, such improvements did not transfer to the validation set, and the overall error magnitude remained comparable. These results suggest that the instability is not specific to any particular architectural choice or hyperparameter configuration.

In addition to quantitative evaluation, we conducted a qualitative analysis through visualization of event frames. By comparing event frames sampled from different temporal windows within the same stimulus, it can be observed, as illustrated in the figure 6.2, that even under the same gaze label, the accumulated event frames differ noticeably in spatial structure. In the example corresponding to the highlighted yellow screen coordinate, the event frame contains not only information from the stable fixation phase but also event signals generated during the eye movement from the previous target to the current one.

This observation indicates that, under the current task design, a stable and consistent supervisory relationship between model inputs and output labels is difficult to establish. Because the precise onset and offset times of eye movements are not explicitly annotated for each participant and each stimulus, a single gaze label often corresponds to multiple distinct eye movement phases. As a result, the input data associated with the same supervision label do not share a consistent physical meaning

across time.

Furthermore, within the same labeled interval, events are primarily generated during eye movement phases, whereas significantly fewer events are produced during stable fixation, and in some extended fixation periods, almost no eye-related events are recorded. Consequently, even if one attempts to retain only the latter portion of each stimulus to enhance supervisory stability, the effective number of informative events becomes insufficient to construct a sufficiently large and consistent training set.

Taken together, these observations characterize the empirical behavior of the gaze prediction task under the current data acquisition paradigm and supervision design: validation errors remain consistently high across model configurations; the training pipeline functions correctly under controlled conditions; and the event patterns associated with a single gaze label exhibit substantial temporal variability. These findings provide a basis for the subsequent analysis of task feasibility and supervision design considerations.

6.3.2 Cognitive Load Classification Analysis

Experimental results reveal a clear performance gap between the within-subject and cross-subject settings for cognitive load classification. Under the within-subject training setting, the model achieves consistently high classification accuracy for the vast majority of subjects, with accuracies stably exceeding 0.95. Only a small number of subjects (e.g., Subject 3 and Subject 10) exhibit relatively lower performance. In contrast, under the cross-subject leave-one-subject-out (LOSO) setting, overall classification accuracy drops substantially to approximately 0.39, accompanied by large variability across test subjects. These results indicate that while the model can effectively exploit subject-specific discriminative cues under within-subject conditions, such cues do not exhibit stable consistency across different individuals.

This observation is consistent with findings from prior studies in cognitive science and eye-movement research [15, 12], which report that working memory load systematically modulates oculomotor dynamics. These studies demonstrate that load-related effects can be detected in measurable eye-movement statistics, including fixation behavior and temporal patterns, supporting the view that cognitive state variations are reflected at the level of observable oculomotor signals. Previous work has shown that the modulation of eye movement behavior by memory load varies significantly across individuals. Under identical load conditions, different subjects may exhibit distinct patterns in fixation stability, blinking behavior, and micro-movement dynamics. As

a result, memory load does not correspond to a single, subject-invariant behavioral representation, but instead manifests as subject-specific modulation patterns.

Additional evidence is provided by the confounding cue analysis presented in Section 4.2. That analysis demonstrates that low-level statistical indicators, such as event rate, do not show stable and consistent responses to condition changes even at the within-subject scale. Instead, their fluctuation patterns are more characteristic of individual-specific variation rather than effects uniformly driven by experimental conditions. Consequently, under cross-subject settings, models cannot reliably construct consistent decision boundaries based on these low-level statistical cues.

Furthermore, to examine whether memory load differences are reflected in low-level event statistics, we introduce a linear support vector machine (SVM) classifier that operates solely on aggregated event-level features. The purpose of this experiment is to assess whether, under a within-subject setting, memory load can be distinguished based purely on the statistical properties of the event stream. Importantly, this analysis is not intended to develop a competitive classification model, but rather to evaluate the intrinsic discriminative information contained in the underlying event statistics when subject identity is held constant.

In this experiment, a fixed number of events (10,000) is extracted from the raw event stream for each sample, from which a set of low-dimensional statistical features is computed. These features cover polarity-related event counts, event rate and its temporal variability, as well as coarse statistics describing the overall spatial distribution of events. A within-subject evaluation protocol is adopted, in which the data of each subject are independently split into training and testing sets.

Due to the experimental design, the slow conditions correspond to longer temporal windows and therefore produce a larger number of samples. If overall accuracy were used directly, the evaluation results would be inevitably biased by class imbalance. To address this issue, balanced accuracy is adopted as the primary evaluation metric, ensuring equal weighting across different memory load levels and presentation speed conditions.

The experimental results are shown in Fig. 6.1. Overall, the within-subject balanced accuracy is mainly concentrated averages around 61%, which is clearly above the random baseline of a six-class classification task (16.7%). This result indicates that, under fixed-subject conditions, memory load and temporal rhythm can already be discriminated to a certain extent using only low-level event statistical features. This observation suggests that memory-load-related information is not confined to high-level semantic or complex spatiotemporal representations, but is already partially reflected

in the statistical distribution of event data.

Further analysis of the feature weights learned by the linear SVM, as illustrated in Fig. 6.3, shows that the classification decisions are primarily driven by polarity-related event count features as well as event rate and its variability, whereas spatial distribution statistics contribute relatively little. This finding indicates that, within a single subject, the effect of memory load on eye movement behavior is mainly manifested through changes in event generation intensity and temporal structure, rather than through pronounced alterations in the overall spatial distribution of events. In other words, even without explicitly modeling fine-grained spatial structure, it is still possible to capture memory-load-induced behavioral differences to a certain extent by characterizing the statistical distribution of events.

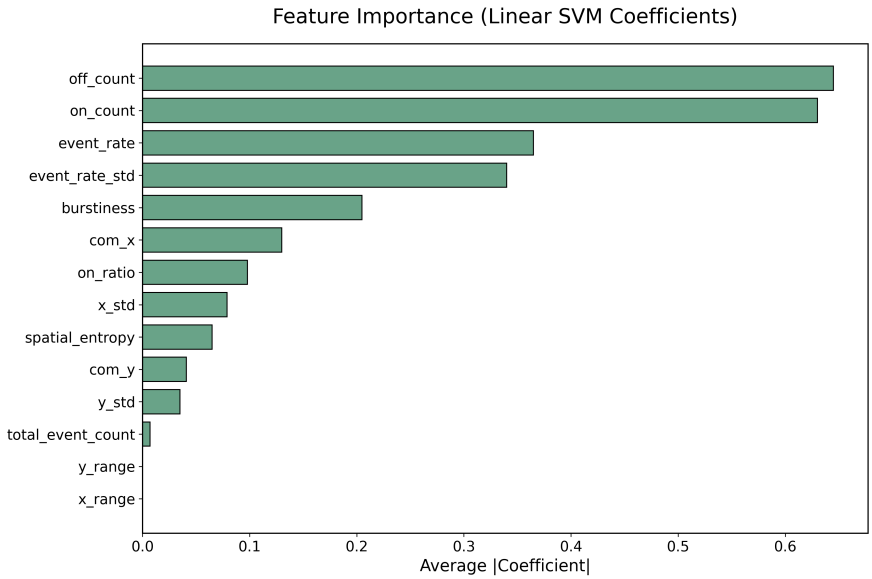

Figure 6.3: Linear SVM feature importance (average —coefficient—). Event count and rate-related features dominate classification, whereas spatial statistics have limited contribution, indicating that memory load effects are mainly expressed through event intensity and temporal structure.

Taken together, this analysis demonstrates that, under within-subject conditions, memory load information is already partially encoded in low-level event statistical features. This result provides a useful baseline for the design of more complex models, while also highlighting that reliance on low-level statistics alone is unlikely to support stable discrimination across all subjects and conditions. Consequently, these findings motivate the exploration of higher-level spatiotemporal modeling and cross-subject generalization analysis in subsequent work.

Overall, the observed high classification performance under within-subject training

settings can be largely attributed to subject-specific statistical regularities present in the event data. When such regularities are not consistent across different subjects, cross-subject classification performance degrades accordingly. This result reveals a core challenge faced by event-camera-based cognitive state recognition, namely how to disentangle subject-specific low-level statistical cues from condition-related information that is genuinely transferable across individuals.

Chapter 7

Discussion

This chapter discusses the implications of the experimental results. We analyze the structural factors that affect model stability and generalization. These include supervision design, characteristics of event-based representations, and inter-subject variability. The aim is to clarify the applicability boundaries of the current modeling paradigm and to identify directions for future improvement.

7.1 Analysis of Gaze Point Prediction under the Current Dataset Setting

The instability observed in gaze regression arises from a mismatch between the event-based input representation and the supervision design under the current dataset setting.

In this work, gaze prediction is formulated as a regression problem. Each single-frame event representation is required to predict a precise screen-space gaze coordinate. However, gaze labels are derived from discrete fixation targets. There is no explicit annotation of saccade onset, offset, or stable fixation phases. As a result, the same gaze coordinate may correspond to different eye movement states at different times.

From the perspective of event sensing, these states produce substantially different event patterns. Saccades generate dense and spatially distributed events. Stable fixation produces sparse and near-static frames. Therefore, even when the supervision label remains unchanged, the associated event frame may encode very different motion information. This weakens the consistency between input and label and makes stable

regression difficult.

Label interpolation does not fundamentally resolve this issue. Although interpolation creates a continuous trajectory between discrete targets, it does not guarantee that each coordinate corresponds to a consistent eye movement state. The same coordinate may still appear under different dynamic conditions. This limits the model's ability to learn a reliable mapping.

Under the present setting—single-frame event input with fixation-based supervision—screen-space gaze regression does not yield stable generalization. This outcome reflects constraints of the dataset design rather than a fundamental limitation of event cameras. Tasks that align more closely with the dynamic sensing characteristics of event cameras, such as eye movement phase recognition or relative gaze change estimation, are structurally more compatible under the current acquisition paradigm.

7.2 Within-Subject and Cross-Subject Performance in Memory Load Classification

In contrast to gaze prediction, memory load classification exhibits clearer but condition-dependent performance characteristics.

Under a within-subject training setting, the model can reliably distinguish different memory load conditions. Under a cross-subject leave-one-subject-out (LOSO) setting, overall classification performance drops substantially. Performance also varies considerably across test subjects. This contrast suggests that memory-load-related patterns are internally consistent within individuals but do not transfer reliably across subjects.

Prior studies in cognitive load and eye-movement research report substantial inter-individual differences in behavioral responses under identical task demands. Under the present acquisition paradigm, these individual differences appear in the statistical structure of the event stream. As a result, decision boundaries learned from one subject cannot be directly reused for another.

The performance gap between within-subject and cross-subject settings reflects subject-dependent modulation patterns embedded in event statistics. This observation motivates a closer examination of the specific information sources exploited by the model and their relationship to low-level statistical features.

7.3 Role of Low-Level Event Statistics in Memory Load Discrimination

To further understand the asymmetry between training settings, we examine whether low-level event statistics already encode discriminative information under fixed-subject conditions.

Statistical analysis shows that event counts, event rates, and their temporal fluctuations exhibit clear individual-specific patterns. Based on this observation, a linear support vector machine (SVM) using only low-level statistical features was employed as an analytical tool. The purpose of this experiment is not to construct a competitive classifier. Instead, it is designed to evaluate whether these features alone are sufficient for within-subject memory load discrimination.

Results indicate that, for most subjects, SVMs based solely on low-level statistics achieve accuracies well above chance. Feature-level analysis shows that discrimination primarily depends on event quantity and temporal characteristics. These include polarity-specific counts, event rates, and their variability. In contrast, coarse spatial statistics derived from event locations contribute comparatively little under the same conditions.

These findings suggest that, for a fixed subject, memory load effects are more strongly reflected in changes in event intensity and temporal dynamics than in stable spatial structure. However, these statistical patterns are not consistent across individuals. Consequently, decision boundaries formed from such features remain stable within subjects but do not generalize across subjects.

Low-level event statistics therefore provide sufficient information for within-subject memory load discrimination. At the same time, their individual-dependent nature explains the limited cross-subject generalization observed in deep models.

7.4 Implications of Temporal Event Statistics for Fast State Discrimination

The effectiveness of low-level statistical features suggests that, for tasks driven primarily by temporal behavioral modulation, detailed modeling of full spatial structure may not always be necessary.

In the present setting, discriminative information is largely reflected in event frequency, intensity, and temporal fluctuation patterns. Stable spatial configurations

play a smaller role.

Because event cameras respond directly to brightness changes, they naturally capture state-dependent variations in temporal dynamics. Meaningful discrimination can therefore be achieved using statistical summaries of event activity. This does not require dense frame-based representations.

This property is advantageous in scenarios with strict latency or computational constraints, such as rapid cognitive state detection or lightweight monitoring systems.

This implication does not suggest that spatial modeling is unnecessary in general. Instead, it indicates that when task objectives are closely tied to temporal modulation, event-based statistical representations provide an efficient and structurally aligned modeling strategy.

7.5 Limitations and Future Directions

Despite achieving stable within-subject performance in memory load classification and providing a systematic analysis of event statistics, this study has several limitations related to dataset design and task formulation.

First, event-based gaze prediction is structurally constrained under the current acquisition protocol. Reliable gaze regression requires continuous trajectory recording, precise temporal alignment, and stable eye–screen geometry. These conditions are not fully satisfied in the present dataset. Consequently, performance is influenced by event sparsity, geometric variation across subjects, and label uncertainty. This limitation reflects characteristics of the dataset rather than an inherent weakness of event cameras.

Second, the modeling paradigm adopted here is particularly effective for tasks centered on temporal behavioral modulation. While event statistics provide informative signals for cognitive state recognition without precise spatial alignment, the current formulation is less suitable for high-precision spatial regression.

Future work aiming to jointly support gaze estimation and memory load recognition would benefit from additional constraints or auxiliary information. For gaze prediction, improvements may include continuous gaze trajectory recording, explicit saccade annotation, or subject-specific geometric calibration. Integration with complementary modalities such as frame-based imagery, head pose estimation, or eye geometry modeling may further enhance spatial robustness. Subject-adaptive modeling or lightweight calibration strategies may also help mitigate inter-subject variability.

In summary, this study highlights both the strengths and structural constraints

of event-based eye movement modeling under the present dataset design. Temporal statistical features enable effective within-subject cognitive state recognition. Stable spatial gaze regression requires more precise supervision and acquisition alignment. These findings provide guidance for future dataset construction and model formulation in event-driven eye-tracking research.

Chapter 8

Conclusion

This work introduced EveLoad, an event-based eye-tracking dataset that jointly provides eye event streams, gaze-point annotations, and graded cognitive workload labels under a task-driven and gaze-constrained experimental paradigm. By integrating event-based sensing with controlled N-back manipulation, the dataset establishes a unified benchmark for studying the interaction between eye dynamics, spatial gaze estimation, and cognitive state recognition.

Unlike existing eye-tracking datasets that primarily emphasize geometric gaze accuracy or rely on free-viewing paradigms, EveLoad adopts a full-screen, randomly guided target-following task that explicitly constrains spatial gaze dispersion. This design reduces reliance on coarse spatial cues and enables systematic investigation of fine-grained temporal eye dynamics under varying workload conditions. The resulting dataset supports two complementary benchmark tasks—gaze-point prediction and memory-load classification—within a shared sensing and annotation framework.

Through structured data-quality validation and confounding-cue analysis, we demonstrated that the collected event streams are temporally continuous, properly aligned with supervision signals, and free from trivial condition leakage via simple sensory statistics. Based on this validated foundation, we evaluated multiple baseline and lightweight models, characterized attainable performance ranges, and analyzed failure modes under both within-subject and cross-subject settings. The results reveal a structural contrast: while temporal event statistics enable effective within-subject cognitive load discrimination, cross-subject generalization remains limited, and high-precision gaze regression is constrained by acquisition geometry and supervision granularity.

Taken together, these findings suggest that event-based eye signals encode informa-

tive temporal modulation patterns that are particularly suited for dynamic cognitive state recognition. At the same time, robust spatial gaze estimation under task-driven constraints requires stronger geometric supervision and alignment mechanisms. By releasing EveLoad together with benchmark protocols and analysis pipelines, this work provides a reproducible experimental basis for future research on event-driven eye tracking, reliability-aware modeling, and cognitively adaptive interaction systems.

Bibliography

[1] Anastasios N. Angelopoulos et al. "Event-Based, Near-Eye Gaze Tracking Beyond 10,000 Hz". In: *arXiv preprint arXiv:2004.03577* (2020). DOI: 10.48550/arXiv.2004.03577. URL: https://doi.org/10.48550/arXiv.2004.03577.

[2] Karlene Ball et al. "Age and Visual Search: Expanding the Useful Field of View". In: *Journal of the Optical Society of America A* 5.12 (1988), pp. 2210–2219. DOI: 10.1364/JOSAA.5.002216. URL: https://doi.org/10.1364/JOSAA.5.002216.

[3] Frédéric Dehais et al. "Physiological and subjective evaluation of mental workload using eye-tracking". In: *Applied Ergonomics* 73 (2018), pp. 242–252. DOI: 10.1016/j.apergo.2018.06.013. URL: https://doi.org/10.1016/j.apergo.2018.06.013.

[4] Brian Guenter et al. "Foveated Rendering". In: *ACM Transactions on Graphics (TOG)* 31.6 (2012). DOI: 10.1145/2366145.2366183. URL: https://doi.org/10.1145/2366145.2366183.

[5] Susanne M. Jaeggi et al. "The concurrent validity of the N-back task as a working memory measure". In: *Memory* 18.4 (2010), pp. 394–412. DOI: 10.1080/09658211003702171. URL: https://doi.org/10.1080/09658211003702171.

[6] Nilli Lavie. "Perceptual Load as a Necessary Condition for Selective Attention". In: *Journal of Experimental Psychology: Human Perception and Performance* 21.3 (1995), pp. 451–468. DOI: 10.1037/0096-1523.21.3.451. URL: https://doi.org/10.1037/0096-1523.21.3.451.

[7] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. "A 128×128 120 dB 15 µs Latency Asynchronous Temporal Contrast Vision Sensor". In: *IEEE Journal of Solid-State Circuits* 43.2 (2008), pp. 566–576. DOI: 10.1109/JSSC.2007.914337. URL: https://doi.org/10.1109/JSSC.2007.914337.

[8] Wenxuan Liu et al. "FovealNet: Advancing AI-Driven Gaze Tracking Solutions for Efficient Foveated Rendering in Virtual Reality". In: *IEEE Transactions on Visualization and Computer Graphics* 31.5 (2025), pp. 3183–3193. DOI: 10.1109/TVCG.2025.3549577. URL: https://doi.org/10.1109/TVCG.2025.3549577.

[9] Anjul Patney et al. "Towards Foveated Rendering for Gaze-Tracked Virtual Reality". In: *ACM Transactions on Graphics (TOG)* 35.6 (2016), pp. 1–12. DOI: 10.1145/2980179.2980246. URL: https://doi.org/10.1145/2980179.2980246.

[10] Alexander Plopski et al. "The Eye in Extended Reality (XR): A Survey on Gaze Interaction and Eye Tracking". In: *ACM Computing Surveys* 55.3 (2022), pp. 1–39. DOI: 10.1145/3528221. URL: https://doi.org/10.1145/3528221.

[11] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. "Peripheral vision and pattern recognition: A review". In: *Journal of Vision* 11.5 (2011), p. 13. DOI: 10.1167/11.5.13. URL: https://doi.org/10.1167/11.5.13.

[12] Kerri Walter and Peter Bex. "Cognitive load influences oculomotor behavior in natural scenes". In: *Scientific Reports* 11 (2021), p. 12405. DOI: 10.1038/s41598-021-91845-5. URL: https://doi.org/10.1038/s41598-021-91845-5.

[13] Pauline van der Wel and Henk van Steenbergen. "Pupil dilation as an index of effort in cognitive control tasks: A review". In: *Psychonomic Bulletin & Review* 25.6 (2018), pp. 2005–2015. DOI: 10.3758/s13423-018-1432-y. URL: https://doi.org/10.3758/s13423-018-1432-y.

[14] Christopher D. Wickens. "Multiple Resources and Mental Workload". In: *Human Factors* 50.3 (2008), pp. 449–455. DOI: 10.1518/001872008X288394. URL: https://doi.org/10.1518/001872008X288394.

[15] Nicholas J. Wyche, Mark Edwards, and Stephanie C. Goodhew. "An updating-based working memory load alters the dynamics of eye movements but not their spatial extent during free viewing of natural scenes". In: *Attention, Perception, Psychophysics* 86 (2024), pp. 503–524. DOI: 10.3758/s13414-023-02741-1. URL: https://doi.org/10.3758/s13414-023-02741-1.

[16] Mark S. Young et al. "State of science: mental workload in ergonomics". In: *Ergonomics* 58.1 (2015), pp. 1–17. DOI: 10.1080/00140139.2014.956151. URL: https://doi.org/10.1080/00140139.2014.956151.

[17] Guoqi Zhao et al. "EV-Eye: Rethinking High-Frequency Eye Tracking through the Lenses of Event Cameras". In: *Advances in Neural Information Processing Systems (NeurIPS 2023)*. Vol. 36. Curran Associates, Inc., 2023. URL: https://papers.nips.cc/paper_files/paper/2023/hash/c41b5d8c1ba15b2aa83e4fa1541f02c8-Abstract-Datasets_and_Benchmarks.html.