NextGen Scientific Publication
American International Journal of Computer Science and Technology

Volume 7 Issue 5, Pg. No. 133-144, AIJCST-V7I5P112, 2025
https://doi.org/10.63282/3117-5481/AIJCST-V7I5P112

Original Article

A Failure Taxonomy of On-Device AI in Head-Worn Displays

***Krishnan Visweswaran**
Independent Researcher, USA.

Article History:

Received: 10.08.2025

Revised: 11.09.2025

Accepted: 22.09.2025

Published: 29.09.2025

Abstract:

Incorporating on-device artificial intelligence (AI) into head-worn displays is becoming commonplace to aid real-time perception, contextual interpretation, augmented reality and intelligent user interaction. Implementing AI models on resource-constrained wearable devices comes with reliability challenges, including sensor quality, environmental factors, model inference, multimodal data fusion, computational constraints, thermal management, connectivity and user safety. In this paper, a systematic failure taxonomy of on-device AI in HWD is presented, where each failure category belongs to one of the following: Sensing, Perception, Inference, Contextual Reasoning, Runtime Execution, Edge–Cloud Coordination, Human-AI Interaction, Privacy and Safety. The taxonomy explores both the root causes and the symptoms of these failures, along with dependencies and implications across layers of the AI processing pipeline, to better understand the effects of individual failures on performance and user experience. In addition, it highlights central evaluation points, such as the inference latency, prediction accuracy, use of resources, energy consumption, the severity of the failure, and the ability to recover from the failure, for consistent failure characterization and comparative evaluation. The proposed taxonomy offers a structured framework for researchers and developers to detect reliability gaps, create systematic evaluation evaluation strategies, and design reliable, resource-efficient, and safety-oriented on-device AI systems for future generation head-worn displays.

Keywords:

Head-Worn Displays, Augmented Reality, Failure Taxonomy, Edge AI, Multimodal Perception, Model Reliability, Resource Constraints, AI Safety, On-Device Artificial Intelligence.

1. Introduction

1.1. Background and Motivation

Head-worn displays are becoming significant wearable computing platforms which incorporate augmented reality, computer vision, contextual assistance, and intelligent human–computer interaction. By incorporating artificial intelligence (AI), these devices can analyse visual scenes, identify objects, comprehend user activities, and provide relevant information in real-time. [1] The head-worn displays are different from traditional computing systems in that they must meet many tight constraints such as battery capacity, processing power, thermal management, device weight and user comfort. These constrains complicate the operation of the reliable AI, especially in conditions of dynamic environment with varying illumination, rapid head movements, partial occlusion and unpredictable user activities. Therefore, it is crucial to identify and classify failures of AI for that to be understood and to enhance the reliability of system operation and to assist in the safe deployment of intelligent head-worn displays.

Copyright by the Author(s). This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-NC-ND 4.0) License (https://creativecommons.org/licenses/by-sa/4.0/)

1.2. On-Device AI in Head-Worn Displays

On-device AI allows head-worn displays to perform machine learning inference on an embedded processor, mobile system-on-chip or dedicated neural processing unit without relying on continuous connectivity to a remote cloud service. This architecture can minimize the latency of communications, enhance responsiveness, maintain connectivity independence and enable increased control of sensitive sensor information. On-device AI applications encompass object detection, scene understanding, visual tracking, speech recognition, gesture recognition, and multimodal context reasoning. [2] These workloads bring challenges to the local execution, however, including interdependent aspects of model complexity, memory availability, processing throughput, energy usage, and thermal management. Moreover, the reliability of the output from AI is dependent on the reliability of the sensors, the environment, the need for generalization of the model, and the synchronization of several data streams. The knowledge of these interactions underlies a systematic taxonomy of the different types of failures in head-worn AI systems for sensing, perception, resource-management, and user-interaction failures.

1.3. Challenges in Reliable AI Operation

For the AI to function reliably while the user is wearing it, the system must be robust, stable, and perform well under varying environments, workloads, and user actions. The quality of input may be affected by sensor noise, motion blur, poor lighting conditions, and occlusion, while unfamiliar scenes and distribution shifts may lead to wrong predictions or low confidence scores. At the same time, the limited memory, battery life, thermal throttling, and scheduling delays may disrupt inference or hinder AI output to meet real-time needs. Errors may also cascade through the layers of the system, with wrong perception leading to false contextual reasoning, and the user being shown wrong information. In safety-critical contexts, these failures can result in loss of situational awareness, task failure or operational risk to the users. To evaluate and build robust on-device AI systems with limited resources and safety awareness, a comprehensive failure taxonomy should include failure type description, failure root causes, observable failure symptoms, failure severity, propagation pathway, and failure impact.

2. Background and Theoretical Foundations

2.1. Architecture of Head-Worn Display Systems

A head-worn display system combines optical displays, cameras, inertial measurement devices, microphones, embedded processors, memory, wireless interfaces and power management devices to enable interactive wearable computing. [3] Each of these modules creates a networked architecture in which sensors collect information about the environment and from the user, processing units perform AI tasks and display modules provide context based on what they see. Accurate sensor synchronization, efficient data transfer, stable processing and timely delivery of visual or auditory feedback are key to the reliability of the overall system. Failure can occur at various levels of the architecture due to hardware constraints, disturbances in the environment, sensor misalignments, and communication delays, and these errors can impact later AI functions. Hence, the importance of grasping the system architecture in order to locate the point of failure, understand the dependency between the components, and categorize the failure based on its location and effect in the head-worn AI processing flow.

2.2. On-Device AI Processing and Edge Intelligence

On-device AI processing involves processing the data and information collected from or on the wearable device, using machine learning or deep learning models, to perform inference without the need for continuous connection to the cloud. [4] By spreading computation over embedded CPUs, GPUs and neural processing units, edge intelligence enables applications like object recognition, scene interpretation, activity detection, and context-aware assistance. The execution of the model, however, must occur under limited memory, energy, processing power and thermal constraints that result in trade-offs between accuracy of prediction, rate of inference and operational efficiency. Reducing resource usage through quantization, pruning, model compression, or adaptive computation can decrease the quality of the prediction or create variations in performance depending on the workload. Therefore, resource depletion, model degradation, inference delays, and interrupted execution are all potential failures that need to be part of a complete taxonomy of failures in on-device AI in head-worn displays.

3. Related Work

Table 1. Comparison of Existing Studies on Wearable and Edge AI Failures

Existing Research Area	Main Focus	Limitation Addressed By This Paper
AI-based perception in head-worn	Object recognition and scene	Limited cross-layer failure classification

displays	understanding	
Edge AI reliability	Model efficiency and resource constraints	Limited integration of hardware and model failures
Embedded system reliability	Hardware faults and runtime stability	Limited focus on AI-specific failures
Multimodal AI	Sensor fusion and contextual reasoning	Limited analysis of failure propagation
Proposed taxonomy	Integrated failure classification	Unified analysis of causes, symptoms, and consequences

3.1. AI-Based Perception in Head-Worn Displays

Previous works on the perception of an HWD have been mainly on object detection, visual tracking, scene understanding, gesture recognition, and context-aware assistance. These features allow for a wearable device to understand the surrounding environment and provide information through AR interfaces. [5] Deep learning models and computer vision technologies have enhanced the accuracy of recognition for different interactive applications, but still need to be adapted to the changes in light, motion blur, partial occlusion, etc., positioning of the camera, and unfamiliar environmental conditions. Studies on perception accuracy and real-time processing have looked at perception failures, but the failure of perception is generally studied in isolation of effects on downstream reasoning, resource constraints, and user impact. It is, therefore, important to have a taxonomy of failures for the system as a whole to link the perception failures with the underlying causes, operational conditions, or propagation across the entire on-device AI pipeline.

3.2. Edge AI Reliability and Model Robustness

The reliability considerations of implementing AI models on resource-limited devices, such as embedded systems, mobile devices, and wearable devices, have been explored in the field of research on edge AI. Current methods analyze model compression, quantization, adaptive inference, hardware acceleration and runtime optimization of the models in an effort to achieve a compromise between prediction accuracy, computational efficiency and energy usage. [6] Other aspects of robustness research include distribution shifts, noisy inputs, adversarial perturbations, and uncertainty in model predictions. Despite these improvements, in many reliability evaluations, only a single aspect of reliability is addressed, such as the reliability of a single performance measure, or a single failure mode.Even with these improvements, it is common practice to focus on reliability of a single aspect of reliability, such as a single performance measure, or a single failure mode, rather than assessing the interactions between the behavior of the model, limitations in the hardware, changes in the environment, and applications requirements. With head-worn displays, these interactions play a special role as they are capable of working collectively to impact real-time responsiveness and user safety. It encourages a cross-layer classification approach which combines with model-level failures runtime and operation failures.

4. Problem Formulation and System Model

4.1. System Model and Operational Assumptions

The proposed system comprises of head mounted display, camera, an inertial sensor, microphone, embedded processors, memory, battery, and visual and audio feedback interface. [7]It is constantly or regularly collecting environmental and user data, and running it through its AI models to make predictions, explain what it sees in context and provide feedback to the user. Assumed operation conditions are constrained computation resources, changing environment, real-time response, and optional edge–cloud communication based upon application design. The system's reliability is based on the availability of sensors, quality of input to the system, execution of the model, allocation of resources available, and timely delivery of output from the system. By making these assumptions, an overall system model is set up for failure identification throughout the various components of the system, including hardware, data acquisition components, AI processing components, and user interaction components, and is not tied to any specific manufacturer or head-worn display platform.

4.2. On-Device AI Processing Pipeline

AI processing pipeline on device includes sensor acquisition, data preprocessing, feature extraction, model inference, contextual interpretation, decision generation, and user facing output delivery. First, cameras and other sensors collect data on the surrounding environment and the actions of the user, then the data are preprocessed before the analysis process by the AI. [8] The machine

learning models then carry out the object recognition, scene understanding, speech processing, multimodal interpretation and so on, while the relevance and presentation of the results generated are determined by the decision logic. Failure can happen at any stage and should cascade to the next stages, e.g. if the quality of the sensor inputs is poor, sensor outputs should not be used for prediction or inference; if inference is delayed by system failures, for example, the resulting visual information should not be used for prediction. Therefore, the pipeline offers a structure to map each failure type to the component it is caused by, the component(s) that will be affected, the symptoms that will be seen, and the business impact resulting from that.

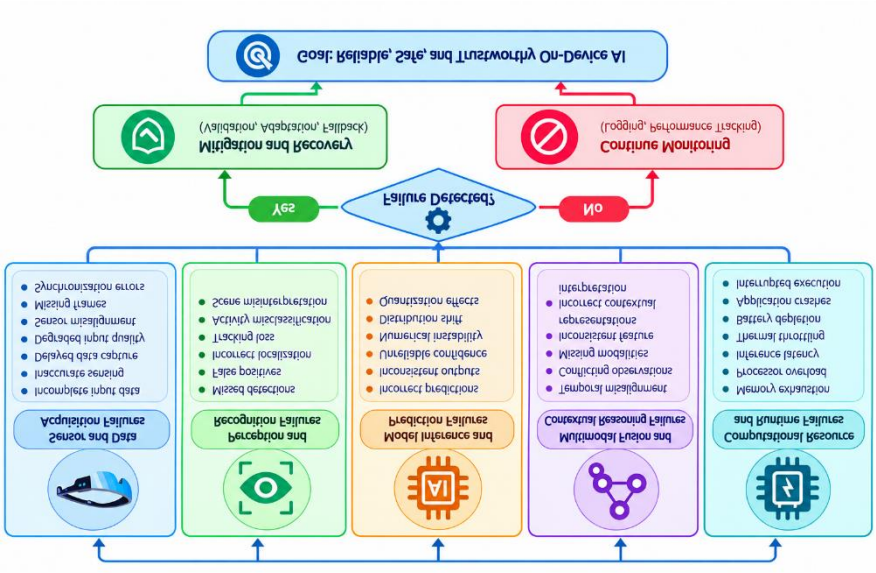


Figure 1. Proposed Failure Taxonomy of On-Device AI in Head-Worn Displays

5. Proposed Failure Taxonomy

5.1. Sensor and Data Acquisition Failures

Sensor and data acquisition failures will happen when the head-worn displays receive the incomplete, inaccurate, delayed, or degraded data needed to process the data on-board. [9] They include poor lighting, camera motion, partial occlusion, sensor noise, camera misalignment, frame loss, and camera/mic/inertial sensor synchronization errors. These conditions may lower the quality of the input, and cause inconsistencies that impact the perception and reasoning steps that follow. The moving speed of the head, for instance, can cause blurring of the images, making it difficult to detect objects and features, and if the sensor streams are delayed, it can lead to associations between objects seen on the screen and the actions of the user, that are incorrect. The identification of these failures, by their source, symptoms and downstream impacts, can lead to the identification of weaknesses in the data acquisition process and the need for ongoing sensor quality monitoring and input validation.

5.2. Perception and Recognition Failures

Perception and recognition failures occur when AI models misinterpret and misrecognize visual scenes, objects, user activities, gestures or environmental events sensed by the head-worn display. [10] These failures have been reported to be missed detections, false-positive classifications, inaccurate object localization, tracking loss, activity misclassification and incorrect scene interpretation. They can be affected by new environments, changes in lighting, occlusion, lack of training data, and differences in the environment at training time and deployment time. In augmented reality applications, perception errors can result in annotations appearing in the wrong location, providing false information that disrupts the user's view of the environment, or potentially even providing false images. To enable AI reliability assessment and enhancement, there has to be a systematic categorisation of perception failures that includes detection, classification, localisation, tracking and scene understanding errors.

5.3. Model Inference and Prediction Failures

Model inference and prediction failures are when the AI model fails to produce accurate, consistent, reliable, or timely results, even though it is receiving input data. [11] Some of these failures can be caused by lack of model generalisation, incorrect confidence

estimation, numerical instability, quantization effects, distribution shifts, etc. or by the mismatch between the assumptions of the model and the actual world. For example, a model trained in indoor settings can have misclassifications in new outdoor settings, and a high-confidence prediction can mask uncertainty for subsequent decision-making logic. These errors can have an impact on contextual assistance, object identification, and real-time decision-making in head-worn displays. Failure characterization should have measures of prediction correctness, model confidence calibration, model output consistency, temporal stability and model behavior under changing conditions to differentiate inference-related failures from sensor acquisition and computation execution failures.

5.4. Multimodal Fusion and Contextual Reasoning Failures

Failure of multimodal fusion and contextual reasoning is when an AI system derives an incorrect conclusion by processing multiple data sources, such as cameras, microphones, inertial sensors and contextual data. Failure can occur due to temporal misalignment, conflicting observations, lack of modalities, different feature representations or misguided assumptions about the relationships between environmental events and user activities. [12] For instance, if there is a mismatch between what the visual information indicates and what the motion data indicates, the system may end up identifying an object with a different type of movement than the visual information, or interpreting what is being said in relation to a visual scene, which can result in providing irrelevant assistance. The failures can cascade beyond the errors in perception to include errors in reasoning that yield conclusions that seem consistent but lack evidence. These failures can be categorized into different types of failures based on different criteria such as synchronization failures, fusion consistency failures, contextual interpretation failures, and evidence reliability failures, to help assess multimodal AI behavior in dynamic wearable environments.

5.5. Computational Resource and Runtime Failures

Table 2. Taxonomy of On-Device AI Failure Categories

Failure Category	Representative Failure Types	Primary Impact
Sensor and data acquisition	Motion blur, sensor noise, missing frames, sensor misalignment	Degraded input quality
Perception and recognition	Missed detections, false positives, misclassification, tracking loss	Incorrect scene interpretation
Model inference and prediction	Incorrect predictions, unreliable confidence, distribution shift, numerical instability	Inaccurate AI outputs
Multimodal fusion and contextual reasoning	Temporal misalignment, conflicting sensor observations, missing modalities, contextual errors	Inconsistent decisions
Computational resource and runtime	Memory exhaustion, processor overload, thermal throttling, application crashes	Delayed or interrupted inference

When the AI workload exceeds the available computational resources and/or the required runtime in the existing hardware and software environment, computational resource and runtime failures occur. Common error scenarios involve running out of memory, overloading the processor, slow inference time, processor overheating, battery depleting, application crashes, and aborted model runs. [13] Such issues are especially problematic on head-worn displays because limited size makes it difficult to achieve cooling, battery size and processing power. Responsiveness can become even worse and outputs may be delayed, incomplete, or unavailable if there is a higher level of complexity in the model or multiple different AI tasks are being executed at the same time. These can also compound perception and reasoning errors as old forecasts are spoon-fed as real-time data. Narrowing resource utilization, execution latency, thermal behavior, energy consumption, and runtime stability analysis, yields a systematic basis to identify these failures, and to design resource-aware scheduling, adaptive inference, and graceful degradation strategies.

6. Failure Propagation and Cross-Layer Dependencies

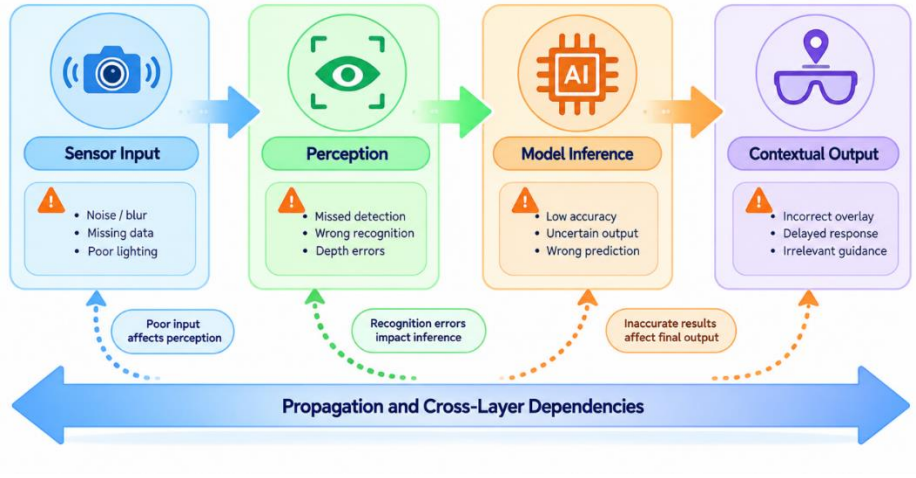


Figure 2. Failure Propagation across the On-Device AI Pipeline

6.1. Failure Propagation across the AI Pipeline

When an error in one phase of the on-device AI system pipeline impacts the next phases of the pipeline, it's called failure propagation and can impact the output of the system. [14] When the display is worn on the head, failures in acquiring the sensor signal can lead to partial or noisy inputs, resulting in a perception error, prediction error and contextual decisions that are incorrect. For instance, when an image is blurred due to motion, the object in the image might be identified incorrectly, causing augmented reality annotations or visual directions to be misplaced. The amount of propagation is dependent on the magnitude of the initial failure, the nature of interdependencies between the processing stages and the efficacy of the error detection mechanisms. Understanding these relationships enables the identification of key failure modes, understanding of local failure modes and system-wide degradation, and the development of reliable methods of monitoring and recovering the AI system from failure at various stages of the processing pipeline.

6.2. Interactions among Sensor, Model, and Runtime Failures

Under the changing environmental and computational conditions, the failures in sensor, model, and run-time can compound each other. The model uncertainty can rise if the quality of the sensor inputs is poor, and computational overload can lead to a delay in inference and to old predictive results. [15] In the same way, if the processing throughput is reduced because of thermal throttling, the users may not get the updates that are needed and may experience not so accurate contextual interpretations when the users move with great speed. The interactions make it hard to determine which failure category is responsible for the degradation of the system without exploring the input quality-model behavior relationship and the runtime conditions. Simultaneous faults, dependency relationships, temporal correlations and contributing factors should therefore be dealt with in a cross-layer failure analysis to elucidate how individual failures interact to affect accuracy, responsiveness, and operational reliability of head-worn AI systems.

6.3. Cascading Failures in Real-Time AI Applications

Cascading failures happen when a first fault leads to a chain of subsequent faults that each exacerbates the issue and ultimately causes a breakdown of the real-time AI application. In head-worn displays, the delayed acquisition of the sensors can delay the object detection, leading to delayed contextual reasoning, inappropriate visual feedback. With repeated attempts to retrieve through further inference or through processing that is heavy on computational resources, computational load may grow and response time may further suffer. [16] This cascading effect is especially relevant in navigation assistance, inspection of industrial machinery, and guidance of hands-free operations, where missed or incorrect information can impact time-sensitive tasks. To identify these failure sequences, developers must study the order in which these events occur, the paths they take, the delays in recovery, and the interdependencies between these systems, so that they can set up early-warning systems, isolate failures and ensure that any isolated errors do not grow into larger application failures.

6.4. Impact on User Experience and Task Completion

AI missteps on the device have direct implications on users, jobs, and trust in on-device head-worn display applications. [17] Misaligned objects, contextually irrelevant recommendations, quick and slow processing, and misrecognition can lead to increased cognitive load, distraction, and loss of usefulness of AI tasks. Recurring failures could also make users ignore system guidance and/or over-rely on potentially incorrect outputs. When the information is misleading or late, or the feedback is late in arriving, it can result in a decrease in situational awareness and therefore increase additional operational risks in safety sensitive environments. The effects of failure are related to frequency, duration, severity, recoverability and importance of failed task. Therefore, the evaluation of failure should not be based solely on technical aspects, like accuracy or latency, but also task completion rate, any interruption in interactions, recovery time, user trust and possible safety implications.

7. Failure Analysis Methodology

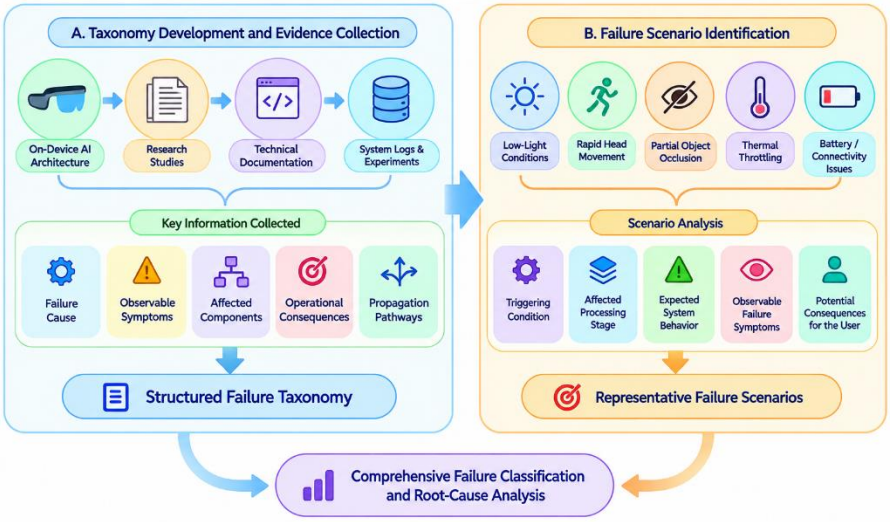


Figure 3. Failure Classifications and Root-Cause Analysis Workflow

7.1. Taxonomy Development and Evidence Collection

The failure taxonomy is constructed by systematically analyzing the on-device AI architectures, head-worn display applications, edge computing constraints and reliability issues mentioned in the literature of relevant research studies. [18] Evidence is laid out in the context of the major processing stages: sensor acquisition, perception, model inference, multimodal reasoning, run time execution, and user interaction. The analysis addresses potential failure causes, visible symptoms, involved system parts, operational impacts and potential propagation paths for each failure identified. Evidence relevant to the problem can be obtained from published sources, technical documentation, system logs, controlled experiments, or documented failure scenarios. This structured approach provides a common base line to organize the various failures into logical groups, differentiate head-worn display related failures from broader edge-AI limitations, and pinpoint research issues that need to be explored or empirically tested.

7.2. Failure Scenario Identification

Failure scenario identification: This process identifies and defines representative scenarios in which on-device AI systems are likely to fail or deliver suboptimal results or become unavailable. [19] Examples of scenarios may be low light conditions, fast head movements, partial occlusion of objects, synchronization errors in sensor data, unknown visual scenes, pressure on memory, thermal control, battery constraints, or intermittent connectivity. Each scenario is characterized by its triggering condition, the processing stage affected, the behavior of the system, what things might go wrong in that system, and what the user might experience as a consequence of those failures. If experimental validation exists, then scenarios can be simulated with controlled tests, variation of workloads, or systematic fault injections; if not, then the scenario can be identified as literature-derived or hypothesized cases. This process ensures full failure coverage and allows the failure behavior to be consistently compared with other head-worn display platforms and AI applications.

8. Experimental Design and Evaluation Scenarios

8.1. Head-Worn Display Test Environment

The head-worn display test environment is designed to test on-device AI reliability in typical indoor-outdoor use scenarios. The evaluation environment needs to vary in illumination, user motions, visibility of objects, background complexity, and availability of sensors to test the impact of environmental factors on the performance of AI. This can be applied to various tasks, such as object recognition, scene understanding, visual tracking, and context-aware assistance. [20] The same measure procedures should be used to record data acquisition, inference outputs, execution latency, resource usage, and observable failure events. Use of a simulation-based or "recorded data" evaluation environment can be used where physical hardware is not available, provided that the limitations are documented. This test environment creates a repeatable context for exploring failure scenarios, testing the behavior of the system under various operating conditions, and determining when on-device AI isn't performing as expected.

8.2. Hardware and On-Device AI Configuration

The evaluation is carried out on the platform that uses the hardware and the on-device AI configuration, which includes the type of hardware, the on-device AI components, the model characteristics, and the runtime software used during the evaluation. [21] These parameters include the type of processor, availability of GPU or neural processing unit, memory, camera resolution, frame rate, battery attributes, OS, and AI inference framework. For the selected models, the model version, the dimensions of the inputs to the model, and the optimization settings for the model should be documented, and the models should represent the target workloads, e.g., object detection, scene classification, or multimodal perception. Evaluation should track the latency of inferences, quality of prediction, memory usage, energy consumption, and temperature behavior under typical workloads. The specification and configuration of hardware, and the conditions of measurement should be reported precisely, and comparisons should be made under consistent conditions of measurement. This setup allows researchers to explore the impact of platform limitations on failure rates, impact, and propagation throughout the on-device AI pipeline.

9. Results and Performance Evaluation

9.1. Distribution of Identified Failure Categories

Table 3. Illustrative Distribution of On-Device AI Failure Categories

Failure category	Failure events (example)	Percentage (%)
Sensor and data acquisition	48	24
Perception and recognition	52	26
Model inference and prediction	40	20
Multimodal fusion and contextual reasoning	32	16
Computational resource and runtime	28	14
Total	200	100

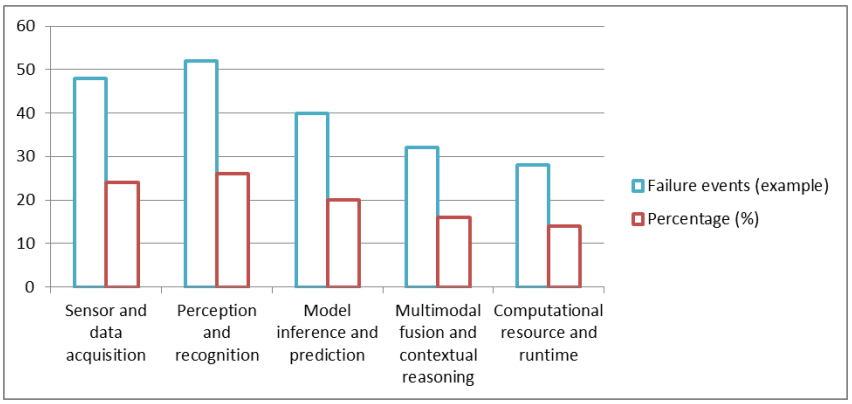


Figure 4. Graphical Illustrative Distribution of On-Device AI Failure Categories

Distribution of identified failure categories gives an overview of the main reliability challenges for on-device AI in head-worn optical displays. [22] Flaws can be categorized as sensor and data acquisition, perception and recognition, model inference, multimodal

reasoning, computational resource management, connectivity, human–AI interaction, and privacy and safety. These categories can be analyzed in terms of the number of failure instances identified, their frequency, and number or percentage of overall system degradation. This analysis can provide information on which processing stages need more attention in the system design and validation. The distribution obtained should be based on the documented literature evidence or actual experimental observation, with the period of data collection, test condition and classification criteria should be clearly stated so that the results can be repeated and understood.

9.2. Failure Frequency across Operating Conditions

Failure frequency under operating conditions assesses the impact of variations in the environment and computation on the reliability of head-worn AI systems. Low illumination, fast head motion, partial occlusion, complex background, higher inference workloads, limited memory, higher temperature of devices, and intermittent connection are all relevant conditions. Failure rates may be determined for each condition as the ratio of failures observed to the number of times that conditions were evaluated. These comparisons can highlight any weaknesses in the environment, weaknesses related to workload, and conditions where there is greater operational instability. Depending on the duration of the scenario, the intensity of the workload and the size of the sample, the analysis should be adjusted accordingly, otherwise it can lead to the wrong conclusions. Information gathered from controlled experiments or systematic observations can be used to aid in the determination of critical operating conditions and the development of more powerful on-device AI systems.

9.3. Failure Detection and Classification Performance

Failure detection and classification measures how well a monitoring mechanism succeeds in detecting abnormal system behavior and correctly classify the observed failures in predefined taxonomy categories. [23] Accuracy, precision, recall, F1 score, false positive rate and false negative rate are some of the relevant evaluation metrics. The overall correctness of the classification is measured by accuracy, whereas precision and recall show the reliability and completeness of the identification of failures. The F1-score is a good compromise to measure precision and recall, especially if there are an unequal number of failures in each category. Evaluations include both category performance and classification results, including differences in performance between categories, and unusual circumstances not witnessed before. When an automated detection mechanism is available, it can be evaluated against the same sets of data and against a consistent set of rules and statistical anomaly detection methods. If there is no actual measurement to support the numerical results, then do not report the numerical results.

9.4. Latency, Memory, Energy, and Thermal Impact

On a resource limited head-worn display, operational effects of AI failures can be measured by latency, memory, energy usage and thermal characteristics. Inference latency is the time taken to run an input through the AI model to produce an output, and memory usage refers to the amount of resources that the AI models and the processes require. [24] Energy measurement can help determine how much battery life is being used by sustained workloads, and thermal sensing can help detect conditions where the workload might cause battery throttling or stall. These metrics must be taken at varying workloads and operating conditions to find correlations between the resource pressure and its failure rate. For instance, a higher model complexity can lead to higher inference latency and energy use, and longer inference time can cause higher device temperature and lower throughput. Average and percentile latency, peak memory, energy per inference and temperature variation give a consistent ground to assess failures risks related to resources and compare deployment configurations.

10. Discussion

10.1. Key Findings and Research Implications

The proposed failure taxonomy offers a structured view of the potential problems with on-device AI in head-worn displays, grouping failures into categories under sensing, perception, inference, multimodal reasoning, computational resources, and user interaction. A major implication is that reliability of AI can't simply be measured by the accuracy of the AI model, as the behavior of the system will be influenced by the quality of the input data, runtime stability, environmental conditions, and timeliness of the output the AI generates. Errors from one operation can travel through to other operations, impacting how the contents are understood within their context and how the content influences user decisions. This cross-layer view underlines the need for integrated failure monitoring, systematic root cause analysis as well as assessment of performance under various operating conditions. The taxonomy can be used for future research to facilitate the comparison of failure scenarios, identify gaps in the understanding of reliability and define a standardized approach to evaluating reliability of wearable AI systems.

10.2. Dominant Failure Categories and Root Causes

The proportion of failure types depends on the application, hardware, environment and AI workload. Failure of the sensor and the perception could be more visible in scenarios with poor illumination, fast motion or even partial occlusion, while runtime failures could become more apparent when the device is subjected to a high amount of computation, a low amount of memory or high temperatures. Model inference failures may be due to distribution shifts, insufficient generalization, or unreliable confidence estimates, and multimodal reasoning failures can occur due to the misalignment of time or conflicting sensor information. These categories may overlap and result in multiple causal factors from the output of the final system. The failure analysis must therefore separate out the cause of failure from contributing causes and consequences, applying coherent system logs, controlled testing and failure classification based on evidence. The dominant categories should ultimately be determined from empirical data and/or systematic literature analysis and not be assumed without evidence.

11. Failure Mitigation and Design Recommendations

11.1. Sensor-Level Validation and Input Quality Control

Therefore, it is crucial to validate the data at the sensor level to stop the poor quality input data from flowing into the on-device AI processing pipeline. To ensure high-quality input for AI inference, head-worn displays should include input-quality checks to detect missing frames, too much motion blur, insufficient illumination, sensor noise, or synchronization issues. The temporal consistency checks and sensor-health monitoring can be used to identify unreliable observations and to assess if further data acquisition is necessary or not. If the input quality is not satisfactory in the predefined thresholds the system can decrease the confidence on future predictions, ask for another sensor data or temporarily stop the unreliable processing. These mechanisms increase the reliability of downstream perception and contextual reasoning and decrease the chances of receiving corrupted inputs to yield incorrect outputs. For wearables, implementing lightweight validation procedures is especially critical as operations on these devices need to be useful and energy-efficient, while also accommodating the limited computational power available.

11.2. Robust Perception and Uncertainty-Aware Inference

Reliable perception and uncertainty-aware inference can enhance the reliability of the system by enabling an AI model to identify a difficult input and convey uncertainty regarding its prediction. Sensitivity to the environment and to scenes that are different from the training set can be mitigated with techniques like using diverse training data, data augmentation, domain adaptation, model calibration, and robustness testing. Inference mechanisms should involve considering the degree of confidence that can be placed in predictions, as well as determining cases where there is not enough evidence to draw reliable inferences. Low-confidence predictions may lead to further checks, alternative runs of models, or revocation of the potentially misleading visual guidance. These are especially significant when using head-worn displays in dynamic environments where the wrong recognition or understanding of the context would impact a user's decision. By integrating strong model building with uncertainty monitoring, prediction errors can have less influence on the resulting operation of the AI on the device, and the resulting uncertainty can serve as a basis for safer and reliable on-device AI operation.

12. Limitations and Future Research Directions

12.1. Limitations of the Proposed Taxonomy

The taxonomy proposed here is a structured categorization of possible failure modes of on-device AI systems for HWD, but its completeness relies on the literature search and the system boundaries as well as the variety of failure modes included. Some failure categories may overlap since there can be sensors degradation, prediction errors from the model, and a lack of computational resources all at once in a failure scenario. In addition, new AI architectures and the rapidly changing wearable hardware could lead to new failure types that are not covered in the existing classification. Validating the taxonomy through controlled experiments and trials in the field could also be uncertain until it is developed using literature analysis, with the caveat that the applicability of the taxonomy to the field remains to be determined. The classification needs to be further improved by empirical evidence, expert assessment, to reduce its coverage, consistency and applicability, but also by the systematic comparison across different applications and also by reproducible testing procedures.

12.2. Hardware and Platform-Specific Dependencies

Failure rates and impacts of on-device AI systems may vary greatly based on hardware architecture, hardware capabilities, sensor characteristics, memory capacity, battery restrictions, thermal design, and software implementation. The CPUs, GPUs, neural processing units, cameras and inertial sensors that are used in different head-worn display platforms could have different performance

requirements and failure modes. Other factors that can affect latency, energy consumption and runtime stability include model optimization techniques, operating system scheduling, firmware behavior, and inference frameworks. Hence, a failure taxonomy for one platform might not completely represent the failure behaviour of another device or application. Future research should expand the proposed classification to other hardware configurations and software environments, as well as determine platform-independent failure patterns and record device-specific dependencies to make it easier to assess the reliability of different platforms and make comparisons.

13. Conclusion

A systematic taxonomy of failure modes for on-device AI in head-worn display systems—reliability issues from sensor acquisition, perception, model inference, multimodal reasoning, computational resources, run-time execution and user interaction. The classification was used to identify how environmental disturbances, limitations of the AI models, hardware restrictions, and operational dependencies can create individual failures, which can result in cascading impacts on the entire AI processing pipeline. The study also identified a systematic approach for failure scenario identification, root cause analysis, operational impacts analysis and failure detection/classification performance evaluation. Additionally, the need for validation of the sensors, uncertainty-aware inference, adaptable resource management, runtime monitoring, and fallback mechanisms to improve the reliability of the system was stressed. While the usefulness of the taxonomy in practice needs to be tested on a variety of devices, workloads and realistic settings, it provides a starting point for a consistent taxonomy of failures and systematic reliability evaluation. The taxonomy should be validated through empirical research and validated evaluation benchmarks should be established, while future on-device AI architecture designs should seek to strike a balance between accuracy, responsiveness, energy efficiency, and user safety in next-generation head-worn displays.

Reference

[1] Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., ... & Adam, H. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1314-1324).

[2] Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017, December). The ML test score: A rubric for ML production readiness and technical debt reduction. In 2017 IEEE international conference on big data (big data) (pp. 1123-1132). IEEE.

[3] Zhou, F., Duh, H. B. L., & Billinghurst, M. (2008, September). Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In 2008 7th IEEE/ACM international symposium on mixed and augmented reality (pp. 193-202). IEEE.

[4] Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., & MacIntyre, B. (2001). Recent advances in augmented reality. IEEE computer graphics and applications, 21(6), 34-47.

[5] Billinghurst, M., Clark, A., & Lee, G. (2015). A survey of augmented reality. Foundations and Trends® in Human—Computer Interaction, 8(2-3), 73-272.

[6] Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E., & Ivkovic, M. (2011). Augmented reality technologies, systems and applications. Multimedia tools and applications, 51(1), 341-377.

[7] Wang, X., Tang, Z., Guo, J., Meng, T., Wang, C., Wang, T., & Jia, W. (2025). Empowering edge intelligence: A comprehensive survey on on-device AI models. ACM Computing Surveys, 57(9), 1-39.

[8] Alam, S., Yakopcic, C., Wu, Q., Barnell, M., Khan, S., & Taha, T. M. (2024). Survey of deep learning accelerators for edge and emerging computing. Electronics, 13(15), 2988.

[9] Musa, A., Kakudi, H. A., Hassan, M., Hamada, M., Umar, U., & Salisu, M. L. (2025). Lightweight deep learning models for edge devices—a survey. International Journal of Computer Information Systems and Industrial Management Applications, 17, 18-18.

[10] Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295-2329.

[11] Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149.

[12] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018, June). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In 2018 IEEE/CVF conference on computer vision and pattern recognition (pp. 2704-2713). IEEE.

[13] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[14] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018, June). Mobilenetv2: Inverted residuals and linear bottlenecks. In 2018 IEEE/CVF conference on computer vision and pattern recognition (pp. 4510-4520). IEEE.

[15] Azuma, R. T. (1997). A survey of augmented reality. Presence: teleoperators & virtual environments, 6(4), 355-385.

[16] Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in neural information processing systems, 28.

[17] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-13).
[18] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in neural information processing systems, 32.
[19] Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136.
[20] Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.
[21] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
[22] Wang, C.-H., Huang, K.-Y., Yao, Y., Chen, J.-C., Shuai, H.-H., & Cheng, W.-H. (2022). Lightweight deep learning: An overview. IEEE Consumer Electronics Magazine, 11(5). https://doi.org/10.1109/MCE.2022.3181759
[23] AI, N. (2023). Artificial intelligence risk management framework (AI RMF 1.0). URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai, 1(0), 0-1.
[24] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[26] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.