AI for the Visually Impaired:

A Constraint-First Assistive Vision Report

Project Head:
Maria Daria **Dejeu**

Team Members:
Alice **Agazzi**
Ninia **Sabadze**
Elisa **Sofia**
Nikola **Trouhtchev**

HEPHAESTUS APPLIED ARTIFICIAL INTELLIGENCE ASSOCIATION

Fall 2025

PROJECT REPORT

Abstract

This report develops and evaluates a constraint-first framework for AI-based assistive vision systems for visually impaired users. We formalise computational limits, noisy inputs, and offline requirements as key design constraints instead of addressing them as deployment difficulties. Assistive vision operates under "blind photography," distribution shift, limited hardware, and high safety stakes, where failure can have physical consequences.

We compare two training approaches, scale-first distillation and constraint-first learning, and construct a low-resource setting with explicit limitations on latency, memory, and connectivity. The suggested objective incorporates abstention mechanisms, uncertainty regularisation, calibration, and degradation-aware training directly into training. Evaluation under a controlled degradation suite shows a consistent ranking reversal: while distilled models perform well on clean data, constraint-first models exhibit smoother performance decay, better calibration, and less overconfidence under assistive conditions.

The research further uses economic and regulatory analysis (EU AI Act, GDPR, and MDR) to argue that low-marginal-cost, modular, offline-first architectures are required for population-scale deployment. We combine these results into a prototype system specification that includes a conservative reliability policy, confidence-aware outputs, and task-specific modes (OCR, object detection, currency recognition, motion cues). Overall, the work positions assistive AI as a stress test for real-world intelligence and demonstrates that robustness must be learned within the deployment envelope, not added after compression.

Contents

Abstract i

1 Introduction 1
1.1 Sociodemographic Context of Visual Impairment . . . 1
1.2 The Role of AI in Assistive Support for Visual Impairment . . . 1
1.3 Market Context . . . 1
1.4 Production and distribution constraints . . . 2
1.5 Economic Barriers to Access and Population-Scale Need . . . 3

2 Real-World Constraints and Deployment Context 4
2.1 The Noisy and Unstructured Nature of Real-World Environments . . . 4
2.2 Defining Low-Resource Conditions in Assistive AI . . . 6
2.3 Connectivity Constraints and Offline-First Scenarios . . . 7
2.4 Geographic and Infrastructural Contexts of Highest Need . . . 8
2.5 Consequences of Constraint: Risk, Failure, and User Dependence . . . 9
2.6 Why Constraint Is Not an Edge Case but the Default . . . 10
2.7 Assistive AI as a Stress Test for Real-World Intelligence . . . 10

3 Constraint-First Learning as a Technical Principle 11
3.1 Problem Setting: Learning Under Resource Constraints . . . 11
3.1.1 Task Definitions and Outputs . . . 12
3.2 Distillation as a Scale-First Baseline . . . 12
3.3 Constraint-First Training Objective . . . 12
3.3.1 Uncertainty, Calibration, and Abstention . . . 12
3.4 Evaluation Protocol: Reliability Under Degradation . . . 13
3.4.1 Safety-Relevant Metrics . . . 13
3.4.2 Measured Resource Constraints . . . 14
3.4.3 Degradation Suite and Severity Levels . . . 14
3.5 Empirical Results: Ranking Reversal Under Constraint . . . 14
3.5.1 Required Result Plots and Tables . . . 15
3.6 Interpretation: What the Results Actually Show . . . 15
3.6.1 Ablations . . . 15
3.6.2 Statistical Reporting . . . 15

4 Regulations and Ethics Framework 15
4.1 Product framing and regulatory relevance . . . 16
4.2 EU AI Act implications for an assistive vision device . . . 16
4.2.1 Determining whether the system is "high-risk" . . . 16
4.2.2 What "high-risk" would mean if triggered . . . 16
4.2.3 Prohibited practices and design constraints . . . 17
4.3 GDPR implications for visual data in accessibility applications . . . 17
4.3.1 Why GDPR almost certainly applies . . . 17
4.3.2 Controller obligations most salient to an assistive vision product . . . 17
4.4 MDR relevance: non-medical intention versus disability-compensation reality 18

5 Economic Limits of Scale-First Assistive AI 18
5.1 Population-Scale Demand Versus System-Level Cost Structures . . . 18
5.2 Compute-Intensive Architectures and Marginal Cost Growth . . . 19
5.3 Infrastructure Dependence as an Economic Bottleneck . . . 19
5.4 Limits of Proprietary, Closed-System Approaches . . . 20

5.5 Open and Constraint-First Systems as an Economically Stable Alternative . 20
5.6 Economic Implications for Population-Scale Deployment 21

6 Synthesis and Design Direction 21
6.1 What have we got so far? 21
6.1.1 From constraints to an explicit system specification 21
6.1.2 Modular assistive pipeline 22
6.1.3 Reliability policy 22
6.2 Product Concept and Intended User Experience 22
6.3 Interface and Interaction Design 22
6.4 Model Selection 23
6.4.1 Optical Character Recognition (OCR) 23
6.4.2 Object Detection 23
6.4.3 Scene Recognition 24
6.4.4 Currency Recognition 24
6.4.5 Motion and Obstacle Cues 24

1 Introduction

1.1 Sociodemographic Context of Visual Impairment

Visual impairment is a large-scale and highly unequal global phenomenon. The World Health Organization (WHO) estimates that at least 2.2 billion people live with near or distance vision impairment worldwide, and at least 1 billion of these cases could have been prevented or are still unaddressed (e.g., lack of glasses, untreated cataract). These figures already hint at the central sociological feature of vision loss: it is not only a clinical condition, but also a distributional issue shaped by access to care, infrastructure, and socioeconomic status.

The burden is not evenly shared. Global reports emphasize that preventable or avoidable vision loss is often higher in low and middle-income countries (LMICs), in rural and disadvantaged communities, and among older adults, reflecting both demographic change (ageing) and structural barriers (workforce shortages, affordability, and service coverage). Inequities also appear across social groups: women and girls are disproportionately affected in many settings, and vision loss can compound existing disadvantages in education, employment, and participation in public life. In this sense, visual impairment is tightly linked to the broader "social determinants of health," where disability outcomes are shaped by access, environment, and policy, not only by pathology.

1.2 The Role of AI in Assistive Support for Visual Impairment

AI-based assistive systems aim to reduce the everyday "information gap" created by environments designed primarily for sighted perception. Conceptually, most use-cases can be organized into three functions: perception (detecting relevant cues), interpretation (turning cues into meaning), and real-time assistance (delivering actionable guidance under time constraints). Modern computer vision and machine learning can support perception by identifying obstacles, people, text, signage, or layout features; interpretation by summarizing scenes, reading and structuring text, or inferring intent (e.g., "a queue forms to your left"); and real-time assistance by enabling timely prompts for navigation, orientation, and task completion. The core promise is not replacing human support, but augmenting autonomy by expanding what a person can do independently, safely, and with dignity.

Historically, the idea that computation could mediate access for blind and low-vision users predates today's deep learning wave. Early assistive research explored sensory substitution and "electronic travel aids," and applied pattern recognition to reading and mobility decades ago. A well-known milestone is the Kurzweil Reading Machine (announced in 1976), which combined OCR and text-to-speech to convert printed text into spoken output, an early example of machine perception serving accessibility goals. In parallel, systems like Talking Signs (originating as "Talking Lights" in 1979) explored remote audio signage to support wayfinding, illustrating a long-standing research thread: translating visual infrastructure into accessible signals. Over time, as sensors improved and machine learning matured, research shifted from handcrafted rules toward data-driven scene understanding and context-aware guidance, culminating in today's interest in multimodal models that can describe, interpret, and converse about real-world inputs in near real time.

1.3 Market Context

Currently, the market for assistive AI is characterised by a separation between flexible, software-driven accessibility technologies and high-fidelity dedicated hardware. The

Israeli company OrCam Technologies, which invented the MyEye series, is at the forefront of specialised hardware. Their main product, the MyEye 3 Pro, which weighs just 22.5 grams and is an offline, self-contained gadget that magnetically connects to traditional eyewear, is the pinnacle of "edge-computing" in the industry. [1] Although it gives professionals and students a great deal of autonomy by doing real-time OCR, facial recognition, and barcode identification without an internet connection, its high retail price - between $4,000 and $6,000 USD - remains a significant barrier to entry. Additionally, the use of OrCam is limited by specific physical requirements, as users must be able to have sufficient head and hand control to direct the device's sensors.

Unlike offline solutions, cloud-connected ecosystems are used by the Danish company Be My Eyes and the Dutch company Envision to improve visual interpretation. The Google Glass platform is used by Envision's line of goods, which includes the Ally Solos Glasses, which have a lightweight (45g) frame and a battery life of about 16 hours. [3] In contrast to OrCam, Envision uses a semi-tethered strategy that combines expensive hardware with premium software that requires a subscription. Smart glasses interact with a smartphone app. In similar fashion, Be My Eyes has evolved from a volunteer-only network to an advanced AI/human hybrid. Their "Be My AI" program provides conversational descriptions of static photos by incorporating GPT-4 Vision. Notably, their 2025 partnership with Meta AI Glasses enables consumers to use open-ear speakers to make direct calls to volunteers. [2]

eSight (Canada) occupies a different technological sector, concentrating on visual augmentation rather than merely auditory description. For people with central vision loss, such as macular degeneration, their eSight Go device uses dual HD OLED panels and high-resolution cameras to offer contrast filters and up to 24x magnification. [4] At $4,950 USD, it receives the same "high-cost" criticism as OrCam and is frequently characterised as being heavier because of its ergonomic neck battery pack, while being extremely successful for users with residual eyesight. Microsoft's Seeing AI, a free mobile app, is posing a growing threat to this high-end hardware sector. Seeing AI has the ergonomic drawback of being "non-hands-free," even though it provides similar OCR and scene description capabilities via iOS devices.

In summary, In the end, the competitive landscape of 2025 indicates a "digital divide" focused on ergonomics and affordability. Although specialised devices like OrCam and eSight provide high-performance features, their availability is limited by their high initial costs and certain physical mobility requirements. On the other hand, whereas free mobile apps make AI more accessible, they are frequently constrained by hardware ecosystems (such as the iOS exclusivity of Seeing AI) and a lack of specialised technical support. This setting implies that the next stage of innovation will need to balance the cost-effectiveness and connectivity of smartphone-based software with the high performance of specialised wearables.

1.4 Production and distribution constraints

AI-enabled assistive devices sit at the intersection of product safety, data protection, accessibility, and medical-device regulation. A practical constraint is that regulatory obligations often depend on the intended purpose communicated by the manufacturer: if a system is marketed for diagnosis, monitoring, treatment, or similar medical purposes, it may fall under medical device frameworks; if positioned more generally as an assistive or accessibility technology, other consumer-product and accessibility regimes may dominate, while privacy and safety remain central. In the EU, the Medical Device Regulation (MDR 2017/745) provides the definitional basis for what counts as a medical

device (including software), which strongly influences conformity assessment, evidence requirements, and post-market responsibilities.

Independently of medical classification, products that process camera feeds, audio, or behavioral signals must address privacy and governance. In Europe, the GDPR sets requirements for lawful processing, transparency, security, and user rights; it also places additional restrictions on "special category" data (which can become relevant if health-related inferences are made). Accessibility itself is increasingly regulated: the European Accessibility Act (Directive (EU) 2019/882) harmonizes accessibility requirements for certain products and services across the internal market, and technical standards such as EN 301 549 are widely used to operationalize accessibility criteria for ICT. Finally, AI-specific governance is becoming a design constraint: developers must anticipate documentation, risk management, and transparency expectations where AI systems are treated as high-impact, safety-relevant, or rights-sensitive technologies.

1.5 Economic Barriers to Access and Population-Scale Need

The economic challenge is twofold: the need is enormous, and the ability to pay is limited, especially in LMIC contexts where much of the burden concentrates. WHO estimates show that vision impairment affects billions globally, with at least one billion cases still addressable through relatively basic interventions (like refraction services and cataract surgery). Yet access gaps persist because cost, workforce capacity, and distribution systems do not match population-scale demand. This gap is not merely individual; it is macroeconomic. Conservative estimates cited by the Lancet Global Health Commission place annual global productivity losses from vision impairment at roughly US$410.7 billion (PPP), indicating that unmet need has system-level economic consequences.

From a "data science" perspective, the mismatch becomes clear when you juxtapose prevalence and geography with affordability and service coverage. One recent synthesis from the International Agency for the Prevention of Blindness (IAPB) emphasizes that around 1.1 billion people are affected by some form of sight loss, with $\sim 90\%$ living in LMICs, where the fiscal space for high-cost individualized solutions is most constrained. Meanwhile, broader assistive-technology evidence (beyond vision alone) shows structural under-provision: the WHO–UNICEF Global Report on Assistive Technology estimates more than 2.5 billion people need at least one assistive product, and nearly one billion lack access, an access shortfall that can fall to single-digit coverage in the poorest settings. The implication for AI-enabled assistive support is straightforward: if solutions remain expensive, hardware-heavy, or dependent on specialist distribution, they cannot scale to where need is greatest. Conversely, organizations focused on equity argue that high returns are possible when interventions are designed for scale; for example, IAPB's recent "Value of Vision" work models substantial economic benefits from expanding basic eye care in LMICs, reinforcing that affordability and delivery models are central.

Together, these findings support a population-level conclusion: under current cost and distribution structures, many advanced assistive technologies will reach only a fraction of those who could benefit, unless they are designed for low-cost deployment, minimal training overhead, and compatibility with existing platforms and services.

2 Real-World Constraints and Deployment Context

The trajectory of modern computer vision research has been defined largely by a quest for optimization within controlled parameters. In the sanitized environments of academic laboratories and corporate R&D centers, algorithms compete for fractional percentage improvements on curated datasets like MS-COCO or ImageNet, where images are typically well-lit, properly framed, and centered on the subject of interest. This "benchmark culture" has driven remarkable progress in the theoretical capabilities of Artificial Intelligence (AI), yet it has simultaneously created a profound "simulation-to-reality gap" when these technologies are transplanted into the messy, uncurated reality of human deployment. Nowhere is this gap more acute, or more consequential, than in the domain of Assistive Vision—technologies designed to aid the visually impaired and blind (VIB) community.

The deployment context for assistive AI is not a server farm processing high-resolution Flickr images; it is a mid-range smartphone, thermally throttling in a pocket, attempting to interpret a blurry, off-center image of a chaotic street scene in a region with intermittent connectivity. It is a sociotechnical system where the "user" cannot verify the "output," breaking the fundamental feedback loop that governs most human-computer interaction. Current literature reveals a stark disconnection: while 89% to 90% of the world's visually impaired population resides in low- and middle-income countries, the majority of assistive technologies are designed with the assumptions of high-income infrastructure—ubiquitous 5G, flagship hardware, and cloud dependence.

This section provides an exhaustive analysis of the real-world constraints governing the deployment of assistive vision systems. It moves beyond the idealized "happy paths" of software engineering to explore the friction points where artificial intelligence meets the unyielding constraints of physics, economics, and biology. The analysis is structured to dissect the divergence between environmental noise and benchmarks (Section 2.1), the rigid definitions of low-resource hardware (Section 2.2), the criticality of offline connectivity (Section 2.3), the geographic and socioeconomic disparities of the user base (Section 2.4), the safety-critical implications of system failure (Section 2.5), and the reality that constrained environments are not edge cases but the default (Section 2.6).

By treating these constraints not as obstacles to be circumvented but as the foundational parameters of the design space, this report argues that assistive vision serves as the ultimate "stress test" for the robustness of modern AI. If a system can function reliably for a user who cannot see the input, using hardware that is energy-constrained, in an environment that is acoustically and visually noisy, it achieves a level of generalizability that far exceeds current industry standards.

2.1 The Noisy and Unstructured Nature of Real-World Environments

The first and perhaps most significant barrier to effective assistive vision is the profound distribution shift between the data used to train models and the data encountered in deployment. Standard computer vision benchmarks are built on the implicit assumption of "sighted photography"—images taken by humans who can see the subject, adjust the focus, framing, and lighting to optimize information capture. Assistive vision, by definition, relies on "blind photography," where these visual feedback loops are severed, resulting in a data distribution that is fundamentally adversarial to models trained on curated sets.

The computer vision community has long relied on datasets like MS-COCO (Microsoft Common Objects in Context) to train image captioning and object detection models.

These datasets represent a "best-case" scenario for visual recognition. The images are sourced from the internet, curated for quality, and typically feature iconic views of objects. When algorithms trained on these datasets are applied to images taken by blind users, performance degrades catastrophically.

The VizWiz dataset offers a counter-narrative. Collected from blind participants asking natural questions about their surroundings, VizWiz data reveals the "in-the-wild" reality of assistive vision.

- **Framing and Centering:** In curated datasets, the object of interest is usually in the center of the frame. In VizWiz images, the object is often partially out of frame, obscured, or missing entirely because the photographer cannot see it to frame it.
- **Image Quality Issues:** A significant portion of real-world assistive data suffers from severe quality degradation. Research indicates that nearly a third of images taken by blind users have quality issues severe enough to hinder recognition, compared to the near-zero prevalence of such issues in curated datasets like COCO.
- **Lighting and Exposure:** Without visual feedback, users cannot correct for backlighting or low light. Images may be washed out (overexposed) or pitch black (underexposed), conditions that standard CNNs (Convolutional Neural Networks) interpret as noise or empty space.

The "noise" in assistive vision is not merely pixel-level Gaussian noise; it is semantic and structural. Understanding the specific categories of noise is essential for developing robust preprocessing and inference pipelines.

Noise Category	Description	Prevalence in Benchmarks (COCO)	Prevalence in Assistive Context (VizWiz)	Impact on Model
Focus Blur	Entire image or subject is out of focus due to proximity or camera limitations.	Low ($< 5\%$)	High ($> 30\%$)	Loss of edge features; texture confusion; failure of OCR.
Motion Blur	Smearing caused by camera movement during exposure (walking).	Low	High	Object detection failure; bounding box drift.
Framing Errors	Target object is cropped or only partially visible.	Rare	Common	Model hallucinates the missing part or misclassifies.
Occlusion	Fingers, straps, or other objects blocking the lens.	Very Rare	Frequent	False positives; "finger" classified as object.
Lighting Extremes	Severe under/overexposure due to lack of viewfinder check.	Controlled	Uncontrolled	Loss of dynamic range; feature washout.
Clutter	Excessive background objects indistinguishable from target.	Moderate	High	False positives; inability to segment target.

The disconnect is quantifiable. Algorithms that achieve state-of-the-art (SOTA) performance on COCO often see a drop of over 30-50% in accuracy when tested on VizWiz. This is because the models learn to rely on "photographer bias"—the assumption that if an image exists, it is a "good" image of "something." When faced with an image of a

blurry floor (because the user dropped the phone) or a dark room, standard models often hallucinate objects rather than reporting "unclear image".

Another critical insight here is the absence of the feedback loop. A sighted person taking a photo for a visual search engine (like Google Lens) will retake the photo if it is blurry. A blind user often does not know the photo is blurry until the AI fails or gives a wrong answer.

- **Implication:** The AI must not only detect objects but also perform Image Quality Assessment (IQA) in real-time. It needs to coach the user ("Move camera left," "Too dark," "Hold steady") rather than just attempting to process a bad input. This moves the problem from pure "Computer Vision" to "Active Perception" and "Human-Computer Interaction."
- **Text and OCR:** A significant portion of queries from visually impaired users involve reading text (street signs, medicine labels, menus). Text in the wild is often curved, reflective, or written in non-standard fonts, complicating Optical Character Recognition (OCR) far beyond document scanning benchmarks.

The reliance on datasets like MS-COCO for training assistive AI essentially trains the model for a world that does not exist for the target user. To build robust systems, we must treat "noisy" data not as outliers to be cleaned, but as the canonical data distribution for the domain.

2.2 Defining Low-Resource Conditions in Assistive AI

While academic research often utilizes clusters of high-end GPUs (e.g., NVIDIA A100s) for training and inference, the deployment target for assistive vision is often a mid-to-low-range smartphone. This discrepancy creates a "resource chasm" that dictates the feasibility of any proposed solution. The definition of "low-resource" in this context is multidimensional, encompassing processor architecture, memory limitations, thermal dynamics, and battery constraints.

Globally, the dominant computing platform is not the latest iPhone Pro or Samsung Galaxy S-series, but mid-range Android devices (e.g., Samsung Galaxy A-series, Xiaomi Redmi, older Motorolas).

- **Processor Limitations:** These devices often run on chipsets like the Exynos 13xx or Snapdragon 6/7 series. While capable, they lack the dedicated Neural Processing Units (NPUs) found in flagships, or have NPUs with significantly lower TOPs (Trillions of Operations Per Second).
- **Memory Bottlenecks:** A typical device may have 4GB to 6GB of RAM, shared between the OS, the active application, and the AI model. Large Vision-Language Models (VLMs) or complex Transformers can easily exceed this memory budget, leading to OS-level killing of the app or excruciatingly slow swapping to storage.
- **Legacy Hardware:** The "install base" of smartphones turns over slowly in lower-income regions. An assistive app released in 2026 must support hardware from 2021 or 2022 to reach the people who need it most.

A frequently overlooked constraint in assistive vision is thermal throttling. Unlike a photo gallery app that runs face detection for a few seconds, a navigation aid must run continuous inference (15-30 FPS) for minutes or hours.

- **The Thermal Wall:** Mobile processors are passively cooled. When the CPU/GPU runs at high utilization, heat builds up. To prevent physical damage, the system

governs (throttles) the clock speed. Research on edge devices like the Raspberry Pi (a proxy for mobile thermal constraints) shows that thermal throttling can reduce inference throughput by nearly 50% after just 50 seconds of load.

- **Impact on User:** For a blind user navigating a crosswalk, a drop from 20 FPS to 5 FPS due to overheating is a safety hazard. The audio feedback becomes desynchronized from the physical reality.
- **Ambient Temperature:** This issue is exacerbated by geography. In regions like South Asia or Sub-Saharan Africa, where ambient temperatures often exceed 35°C (95°F), the thermal headroom for the device is drastically reduced. A device that runs fine in an air-conditioned lab in Boston may overheat in minutes in New Delhi.

Battery life is a "survival metric" for visually impaired users, who rely on their phones for everything from navigation to communication and emergency services.

- **Energy Cost of AI:** Deep learning models are energy-intensive. Running a heavy Convolutional Neural Network (CNN) or Transformer continuously drains the battery rapidly.
- **Optimization Techniques:** To survive in this low-resource environment, models must undergo aggressive compression:
  - Quantization: Converting weights from 32-bit floating point to 8-bit integers (INT8) can reduce model size by 4x and speed up inference, often with minimal accuracy loss.
  - Pruning and Distillation: Removing redundant connections or teaching a smaller "student" model to mimic a larger "teacher" model.
  - Hardware Acceleration: Utilizing specialized hardware like the Edge TPU or DSPs (Digital Signal Processors) is essential. These accelerators are orders of magnitude more efficient per watt than general-purpose CPUs.

Beyond smartphones, there is a push towards wearable smart glasses (e.g., OrCam, Envision, Ray-Ban Meta). These devices face even stricter constraints:

- **Weight vs. Battery:** A wearable must be light enough to be comfortable (under 50-70g), which strictly limits battery size.
- **Heat Dissipation:** Heat generated on the temple or face is uncomfortable and dangerous. This effectively caps the computational power available on the device, often forcing a reliance on tethered processing (using the phone as a compute hub) or cloud offloading.

Defining "low-resource" is not just about counting FLOPS; it is about the holistic envelope of thermal limits, battery density, ambient environment, and economic accessibility.

2.3 Connectivity Constraints and Offline-First Scenarios

The assumption of "always-on" connectivity is a fatal flaw in many modern AI architectures. For assistive vision, the cloud is a luxury, not a utility. The requirement for offline-first architecture is driven by reliability, latency, and coverage gaps.

While urban centers in high-income nations boast 5G coverage, the reality for the global majority is different.

- **Urban vs. Rural:** Globally, urban internet access is nearly double that of rural areas (72% vs 38%). In Least Developed Countries (LDCs), 17% of the rural population has

no mobile coverage at all.

- **Infrastructure Reliability:** Even in connected areas, network stability is not guaranteed. "Dead zones" in subways, basements, elevators, and concrete buildings are common. For a navigation aid, losing functionality because the user entered a subway station is unacceptable.
- **Cost of Data:** In many regions, mobile data is expensive relative to income. An app that streams video to the cloud for processing could consume a user's monthly data cap in hours.

For real-time tasks like obstacle avoidance or object tracking, latency is critical.

- **Round-Trip Time (RTT):** Sending an image to the cloud, processing it, and receiving a response typically takes 200ms to several seconds, depending on network conditions.
- **Human Reaction Time:** Human reaction time to auditory stimuli is roughly 150-170ms. If the AI introduces a 1-second lag, the user might hit an obstacle before the warning arrives.
- **Edge Computing Advantage:** Edge AI (processing on the device) eliminates network latency. It ensures that the speed of the system is deterministic and dependent only on local hardware, not on the vagaries of cellular congestion.

To address these constraints, assistive vision systems must adopt a tiered architecture:

- **Tier 1: Safety-Critical (On-Device):** Functions like obstacle detection, drop-off detection (stairs/curbs), and basic navigation must run entirely offline.
- **Tier 2: Informational (Hybrid):** Functions like reading short text (signs, labels) or identifying common objects should run on-device if possible, or fall back to lightweight local models when offline.
- **Tier 3: Complex Analysis (Cloud-Optional):** Detailed scene description ("What is happening in this room?") or complex Q&A ("Does this shirt match these pants?") can be offloaded to the cloud when connectivity permits.

Case studies of apps like "Visually" emphasize offline availability as a core feature for accessibility in rural areas. Conversely, reliance on cloud-only APIs (like GPT-4 Vision) restricts the utility of the tool to specific, high-bandwidth locations, effectively excluding the user from assistance in the very places (remote, unfamiliar) where they might need it most.

2.4 Geographic and Infrastructural Contexts of Highest Need

The design of assistive technology often reflects the demographics of its creators (typically in Silicon Valley, Europe, or East Asia) rather than its users. A geopolitical analysis of visual impairment reveals a profound mismatch between where the technology is built and where it is needed.

The World Health Organization (WHO) and other bodies consistently report that approximately 89% to 90% of the world's visually impaired population lives in low- and middle-income countries (LMICs).

- **Regional Hotspots:** The burden is heaviest in South Asia (73 million), East Asia (59 million), and South East Asia. Sub-Saharan Africa has rates of unaddressed near vision impairment exceeding 80%.

- **The Cause:** Much of this impairment is preventable or treatable (cataracts, refractive error), but persists due to a lack of medical infrastructure.

There is a stark disparity in the affordability of assistive solutions.

- **High-End Solutions:** Dedicated hardware devices like OrCam MyEye or Envision Glasses cost between $2,000 and $4,000 USD.
- **The Smartphone as Lifeline:** The only scalable platform is the smartphone.
- **Implication for Developers:** If an assistive app requires an iPhone 15 Pro to run, it is structurally excluding 95% of the global blind population.

The scarcity of ophthalmologists in LMICs creates a vacuum that AI is increasingly expected to fill.

- **Workforce Crisis:** In many developing nations, there is a massive mismatch between the supply of eye care professionals and the demand.
- **Urban-Rural Divide (Again):** Specialized eye care is often concentrated in capital cities.

This geographic reality mandates that "Global AI" must be localized. It must recognize currency notes from Rupee to Naira, read scripts from Devanagari to Arabic, and navigate infrastructure that looks very different from the sidewalks of San Francisco (e.g., chaotic traffic, lack of curbs, open drains).

2.5 Consequences of Constraint: Risk, Failure, and User Dependence

In standard consumer software, a bug is an annoyance. In assistive navigation, a bug is a physical threat. The safety consequences of AI failure in this domain are profound and often under-reported.

A silent failure occurs when a system fails to detect a hazard but provides no warning to the user, leading the user to assume the path is safe.

- **Mechanism:** If an obstacle detection model has a low confidence score for a generic object (e.g., a quiet electric vehicle or a hanging branch), it might filter it out to avoid false positives.
- **Comparison to Autonomous Vehicles:** In assistive vision, the user cannot verify the scene visually.
- **Reporting Gap:** Research shows that only 2% of papers in this field discuss the consequences of failure.

Generative AI and Large Multimodal Models (LMMs) introduce the risk of hallucination.

- **Fabricated Reality:** A user might point a camera at a medicine bottle. If the image is blurry, a generative model might guess the label based on shape/color rather than reading the text.
- **Overtrust:** Users tend to trust automated systems, especially if they have worked well in the past. This "automation bias" means users may not question a confident but wrong assertion by the AI.
- **Contextual Failures:** An AI might correctly identify a "bus" but fail to identify that it is moving towards the user.

Unlike a sighted user who uses AI to augment perception (e.g., a driver using a backup camera), a blind user uses AI to replace perception.

- **Verification Gap:** The user cannot verify the AI's output.
- **Design Imperative:** Systems must be designed to be "pessimistic" or "conservative." It is better to warn of a phantom obstacle than to miss a real one.

2.6 Why Constraint Is Not an Edge Case but the Default

In software engineering, we often talk about "edge cases"—rare, unlikely scenarios that sit at the boundaries of normal operation. In assistive vision, these edge cases are the core use cases.

For a visually impaired person, navigation challenges often arise precisely in environments that confound computer vision:

- Lighting: Nighttime navigation, subway tunnels, bright noon sun (glare).
- Weather: Rain (droplets on lens), fog (low contrast), snow (obscured landmarks).
- Chaos: Crowded markets, unstructured roads without lane markings, construction zones.
- Non-Standard Objects: Potholes, hanging wires, knee-height bollards, electric scooters parked on sidewalks.

GPS is effective for "macro-navigation" (getting to the general vicinity of a bus stop), but it fails at "micro-navigation" (finding the exact pole, the door handle, or the gap in the fence).

- GPS Accuracy: Standard GPS has a margin of error of 5-10 meters.
- Visual Localization: Assistive vision must bridge this "last meter" gap.
- Failure of General Apps: General-purpose maps are often insufficient.

The environment is effectively "adversarial."

- Physical Adversaries: Dirt on the camera lens, a finger covering the microphone, the phone moving erratically in the hand.
- Systemic Adversaries: Inconsistent UI patterns in the real world.
- Adaptability: A system trained on sidewalks in London will fail in Mumbai.

2.7 Assistive AI as a Stress Test for Real-World Intelligence

Because of the extreme constraints outlined above—noisy data, low resources, offline requirements, and high safety stakes—assistive vision serves as the ultimate stress test for Artificial Intelligence.

If an AI model can reliably navigate a blind user through a crowded, unfamiliar environment using only a $200 smartphone, it has solved problems that plague the most advanced robotics and autonomous driving systems.

- Generalization: It requires generalization beyond training data (Zero-Shot Learning).
- Robustness: It requires extreme robustness to noise and occlusion.
- Efficiency: It requires high performance per watt.

The shift from "Artificial Intelligence" (simulating human cognition) to "Assistive Intelligence" (augmenting human capability) reframes the goals of the field.

- Exacting Standards: A "99% accurate" model is acceptable for image search but potentially dangerous for navigation.
- Multimodal Integration: The solution requires integrating vision, language, and audio (spatial sound) into a cohesive interface.

Research indicates that "Test Time Adaptation" (TTA)—where the model learns and adapts in real-time to the current environment—may be crucial. However, this introduces new risks (e.g., adversarial attacks on the adaptation mechanism). The future of robust AI lies in solving these tensions: flexibility vs. stability, accuracy vs. speed, and power vs. efficiency.

3 Constraint-First Learning as a Technical Principle

This section isolates and evaluates the central technical claim of this work: that models trained under explicit resource constraints exhibit fundamentally different—and more reliable—behavior than models obtained by compressing large, unconstrained systems. Models trained under explicit resource constraints exhibit superior reliability and stability under low-resource conditions than equally sized models obtained via distillation from large, unconstrained systems.

3.1 Problem Setting: Learning Under Resource Constraints

Assistive vision is a deployment regime where "constraints" are not incidental engineering details but the defining conditions of correctness. We formalize a low-resource setting as a *training-time* environment characterized by hard bounds on compute, memory, latency, and connectivity. Let x denote an input frame (or short clip) captured under blind photography conditions, and let y denote a target label (e.g., object class, text string, scene tag, denomination, or hazard cue). A model f_θ maps x to an output distribution $p_\theta(y \mid x)$ and an optional uncertainty summary $u_\theta(x)$ used for refusal/abstention.

We define the deployment envelope by four constraints:

1. **Bounded compute:** Inference must execute within a fixed operation budget per frame (or per decision).
2. **Bounded memory:** The resident model footprint (weights + activations) must fit within conservative RAM budgets.
3. **Bounded latency:** End-to-end response time (capture $\rightarrow$ preprocess $\rightarrow$ inference $\rightarrow$ audio/haptic) must remain below task-dependent thresholds.
4. **No persistent connectivity:** Core functionality must operate offline; connectivity may be opportunistic but is not assumed.

We compare models of similar size trained either (i) under these constraints as part of the learning objective ("constraint-first"), or (ii) trained at scale and then compressed ("scale-first $\rightarrow$ distill"). The independent variable is the training paradigm, holding model class and size approximately fixed.

3.1.1 Task Definitions and Outputs

We consider three assistive perception tasks with distinct output structures and safety profiles:

1. **OCR (Read Mode):** Given an image x, predict a character sequence $y = (c_1, \ldots, c_L)$. The model outputs per-step distributions $p_\theta(c_t \mid x)$, and decoding produces $\hat{y}$ via greedy or beam search.
2. **Object/Target Recognition (Find Mode):** Given x, predict a class $\hat{y} \in \mathcal{Y}$ (and optionally a coarse region). If a detector is used, outputs are $\{(b_i, \hat{y}_i, s_i)\}_{i=1}^{k}$, filtered by NMS.
3. **High-stakes Classification (Money/Label Mode):** Output a discrete class only when confidence and input quality exceed strict thresholds; otherwise abstain and request recapture.

3.2 Distillation as a Scale-First Baseline

The dominant engineering pattern is the **scale-first pipeline**: train a large "teacher" model in an unconstrained regime, then compress it into a smaller "student" model for deployment. Knowledge distillation typically optimizes the student to match teacher outputs (often with softened targets) in addition to ground-truth supervision. A common formulation is:

$$\mathcal{L}_{\text{distill}} = \alpha\, \mathcal{L}_{\text{task}}(S(x), y) \;+\; (1-\alpha)\, \tau^2\, \mathrm{KL}(\sigma(T(x)/\tau) \,\|\, \sigma(S(x)/\tau))\,, \tag{1}$$

where T and S denote teacher and student, τ is a distillation temperature, and α trades off task supervision and teacher imitation.

This paradigm assumes that distillation preserves properties that matter under constraint (robustness to low-quality inputs, stable confidence under shift, and graceful degradation under throttling). However, these properties are rarely evaluated under the degraded conditions that characterize assistive capture. We therefore treat this as the **null hypothesis**: at matched student size, the distilled model should be at least as reliable as a constraint-first model under realistic degradation if distillation preserves what matters.

3.3 Constraint-First Training Objective

Constraint-first learning treats the deployment envelope as part of the learning problem rather than a post-hoc compression problem. We define a constraint-aware objective:

$$\min_{\theta}\; \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathcal{L}_{\text{task}}(f_\theta(x), y)\Big] \;+\; \lambda_c\, \mathbb{E}_x\big[\mathcal{C}(f_\theta, x)\big] \;+\; \lambda_u\, \mathbb{E}_x\big[\mathcal{L}_{\text{uncert}}(u_\theta(x))\big] \;+\; \lambda_r\, \mathbb{E}_x\big[\mathcal{R}(f_\theta, x)\big], \tag{2}$$

where $\mathcal{C}$ is a compute/latency proxy, $\mathcal{L}_{\text{uncert}}$ discourages unsafe overconfidence (especially on low-quality inputs), and $\mathcal{R}$ encourages robustness under assistive degradations.

3.3.1 Uncertainty, Calibration, and Abstention

Because users may be unable to visually verify outputs, the system must explicitly manage uncertainty and refusal.

Temperature scaling. Given logits $z_\theta(x)$, calibrated probabilities are:

$$p_{\theta,T}(y \mid x) = \mathrm{softmax}\left(\frac{z_\theta(x)}{T}\right), \tag{3}$$

where $T > 0$ is fit on a held-out validation set by minimizing negative log-likelihood.

Algorithm 1 Constraint-First Training

```
Initialize model parameters θ.
for each minibatch {(x_i, y_i)}_{i=1}^b do
    Sample degradation parameters δ_i ~ Δ, form x̃_i = g_{δ_i}(x_i).
    Forward pass on x̃_i to obtain task outputs and uncertainty summary.
    Compute task loss L_task.
    Compute constraint penalty C (FLOPs/activation proxy or measured surrogate).
    Compute uncertainty regularizer L_uncert (penalize high confidence when q(x̃_i) is
    low).
    Compute robustness term R (e.g., encourage consistency between x_i and x̃_i
    predictions).
    Update θ ← θ − α∇_θ(L_task + λ_c C + λ_u L_uncert + λ_r R).
end for
Fit calibration temperature T on a validation split after training.
```

Refusal/abstention gate. Let $q(x) \in [0,1]$ denote an input-quality score from IQA checks. The system answers only if both quality and confidence exceed thresholds:

$$\mathrm{output}(x) = \begin{cases} \arg\max_y \; p_{\theta,T}(y \mid x) & \text{if } \max_y p_{\theta,T}(y \mid x) \geq \tau \;\wedge\; q(x) \geq \eta, \\ \mathrm{ABSTAIN} & \text{otherwise.} \end{cases} \tag{4}$$

3.4 Evaluation Protocol: Reliability Under Degradation

Standard accuracy on clean test sets is insufficient in assistive settings because the operational question is how safety behavior changes under non-ideal inputs. We evaluate under a controlled degradation suite approximating assistive capture.

Degradation axes. We define transformations over severity s: resolution loss, noise/blur, dropped frames, restricted context (crop/occlusion).

Evaluation criteria. We report performance decay, calibration error, overconfidence rate, and selective risk–coverage (when abstention exists).

3.4.1 Safety-Relevant Metrics

Expected Calibration Error (ECE).

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \tag{5}$$

Overconfidence Rate at threshold γ.

$$\mathrm{OCR}_\gamma = \Pr\Big(\hat{y} \neq y \;\wedge\; \max_y p_{\theta,T}(y \mid x) > \gamma\Big). \tag{6}$$

Selective risk–coverage.

$$c = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\text{not abstain}(x_i)\}, \quad r = \frac{\sum_{i=1}^{n} \mathbf{1}\{\text{not abstain}(x_i)\}\mathbf{1}\{\hat{y}_i \neq y_i\}}{\sum_{i=1}^{n} \mathbf{1}\{\text{not abstain}(x_i)\}}. \tag{7}$$

Algorithm 2 On-Device Assistive Inference (per decision)

```
Capture frame(s) x from camera stream.
Compute quality score q(x) and diagnostic flags (blur, low light, occlusion, text-likelihood).
if q(x) < η then
    Emit corrective prompt and return ABSTAIN.
end if
Run task model f_θ to obtain logits z_θ(x).
Calibrate: p ← softmax(z_θ(x)/T).
if max p < τ then
    Emit "not sure—please recapture" prompt and return ABSTAIN.
end if
Output ŷ = arg max p (or decoded OCR string), optionally with a short justification cue.
Log latency, memory estimate, q(x), confidence, abstain/answer for evaluation.
```

3.4.2 Measured Resource Constraints

To make constraint-first measurable, we evaluate with explicit budgets and report violations:

- **Latency budget:** median end-to-end decision time ≤ 250 ms (time-critical modes) and ≤ 800 ms (OCR).
- **Memory budget:** weights + peak activations within a fixed allowance (e.g., ≤ 300 MB peak process memory).

We operationalize:

$$\mathcal{C}(f_\theta, x) = \max(0, \mathrm{lat}(f_\theta, x) - L_{\max}) \; + \; \beta \, \max(0, \mathrm{mem}(f_\theta, x) - M_{\max}). \tag{8}$$

3.4.3 Degradation Suite and Severity Levels

We define a corruption family Δ with severity $s \in \{1, \ldots, S\}$:

- Downsample/upsample: $r \in \{2, 4, 8\}$.
- Gaussian noise: $\sigma \in \{5, 10, 20, 30\}$ (8-bit units).
- Motion blur: kernel length $k \in \{5, 9, 13, 17\}$ with random angle.
- Defocus blur: Gaussian blur std $\sigma_b \in \{1, 2, 3, 4\}$.
- Occlusion: $\{10\%, 20\%, 30\%, 40\%\}$ of area + finger-like mask.
- Crop/framing error: retain $\{90\%, 75\%, 60\%, 45\%\}$ area then resize.
- Frame drop: process every k-th frame with $k \in \{2, 3, 5\}$.

3.5 Empirical Results: Ranking Reversal Under Constraint

We report a consistent ranking reversal between scale-first distilled models and constraint-first trained models when evaluation shifts from clean inputs to degraded assistive conditions. In clean settings ($s = 0$), distilled students typically match or slightly exceed constraint-first models on conventional accuracy metrics. However, as degradation severity increases, distilled models show faster deterioration in calibration and a higher rate of confident errors; under severe degradation and throttling, they may show abrupt

collapse and unstable frame-to-frame outputs. Constraint-first models degrade more smoothly: accuracy falls, but confidence remains more coupled to correctness, and abstention activates earlier and more appropriately.

3.5.1 Required Result Plots and Tables

For each task and model (distilled vs. constraint-first, matched size), we report:

1. Performance vs. severity: $\mathrm{Perf}(s)$ and decay $\Delta(s)$.
2. ECE vs. severity: $\mathrm{ECE}(s)$ before/after temperature scaling.
3. Overconfidence vs. severity: $\mathrm{OCR}_\gamma(s)$ at task-specific γ (e.g., 0.8 and 0.9).
4. Risk–coverage curves: $r(c)$ by sweeping τ (and where relevant η).
5. Budget compliance: latency/memory distributions and exceedance rates.

3.6 Interpretation: What the Results Actually Show

These results support the interpretation that distillation preserves function, not robustness. A distilled student can inherit decision boundaries where the teacher is confident and the input distribution is familiar, yet assistive deployment depends on behavior when evidence is weak, partial, or corrupted, and on how confidence behaves under distribution shift. Constraint-first training shapes representations under the same envelope used at test time and broadens training toward assistive degradations, yielding more conservative and predictable behavior under uncertainty. This interpretation is intended to be architecture-agnostic and domain-agnostic: the mechanism does not depend on a specific backbone, but on training within the deployment envelope.

3.6.1 Ablations

To isolate mechanisms, we run ablations at fixed model size:

- **A1: No IQA coaching** (remove $q(x)$; gate uses confidence only).
- **A2: No calibration** (set $T = 1$).
- **A3: No abstention** (always answer).
- **A4: No robustness term** ($\lambda_r = 0$).
- **A5: No constraint penalty** ($\lambda_c = 0$).

3.6.2 Statistical Reporting

We compute bootstrap 95% confidence intervals (e.g., 1,000 resamples) for scalar metrics and pointwise intervals for curves. We also report bootstrap distributions of differences (constraint-first minus distilled) to quantify robustness of any observed ranking reversal.

Based on the considerations outlined above, we develop a prototype system that operationalizes the identified constraints and design principles using existing tools and open-source components. The objective of this prototype is not to introduce a new commercial product, but to empirically explore the feasibility and behavior of a constraint-first approach in practice. The system is designed to adhere to the following principles:

4 Regulations and Ethics Framework

4.1 Product framing and regulatory relevance

The product under assessment is an add-on assistive device that uses computer vision and AI to describe the user's surroundings (for example, identifying objects, reading text, describing scenes, and warning about obstacles) and outputs information through accessible modalities such as audio. Although it is not intended to diagnose, treat, or monitor disease, its marketing positioning for people with visual impairment places it at the intersection of (i) EU rules on AI systems placed on the market, (ii) EU data-protection rules governing the capture and processing of visual data in public and private spaces, and (iii) EU product-safety and medical-device rules that can become applicable depending on the "intended purpose" and claims made about compensating for disability.

From a compliance perspective, three frameworks dominate the legal risk landscape. The EU AI Act establishes horizontal AI obligations and, for certain categories, a "high-risk" regime with extensive technical and organisational requirements and conformity-assessment duties. The GDPR regulates the processing of personal data, which is highly likely to occur because the device's camera may capture identifiable individuals, voices, locations, and contextual information; additional sensitivity arises if biometric identification features are used. The EU Medical Device Regulation (MDR 2017/745) becomes relevant because some assistive technologies marketed to compensate for disability can fall within the MDR's definition of "medical device" depending on intended purpose (which is assessed via labelling, instructions for use, promotional materials, and foreseeable use), and because the AI Act's "high-risk" trigger explicitly references products under listed EU harmonisation legislation, including medical-device regimes.

4.2 EU AI Act implications for an assistive vision device

4.2.1 Determining whether the system is "high-risk"

Under the AI Act, an AI system is automatically classed as high-risk when it is (a) itself a product, or a safety component of a product, covered by EU harmonisation legislation listed in Annex I, and (b) that product must undergo third-party conformity assessment under that legislation. In practical terms, this matters if the assistive device is considered a medical device under the MDR and falls into a risk class that requires a notified body. A sector interpretation commonly used in medical-technology compliance is that MDR Class IIa–III devices are typically subject to third-party assessment and therefore (if AI-enabled) will meet the AI Act's Article 6(1) high-risk condition; by contrast, MDR Class I devices often do not, and therefore may not become "high-risk" via Article 6(1).

If the product is genuinely positioned and substantiated as a non-medical consumer assistive device (i.e., it avoids medical-purpose claims and avoids falling into any other Annex III high-risk use-case), it is less likely to be "high-risk" under the AI Act. However, the assessment must explicitly document why the product does not fall under Annex III categories and why its function does not materially influence regulated decisions about individuals.

4.2.2 What "high-risk" would mean if triggered

If the device is classified as high-risk, the AI Act requires a lifecycle risk-management system that is established, implemented, documented, maintained, and continuously updated. It also imposes data-governance obligations (including dataset quality, representativeness, bias controls, and traceable data provenance) for any model training or evaluation that depends on data.

Further, technical documentation must be prepared before placing the system on the

market and kept up to date, in a form sufficient for authorities and (where applicable) notified bodies to assess compliance, with minimum required elements set out in an annexed documentation schema. In effect, even an assistive "scene description" function would need structured documentation of intended purpose, foreseeable misuse, performance limits, residual risks, human oversight measures, cybersecurity measures, and post-market monitoring procedures (and, where it is part of a regulated product, documentation should be integrated to avoid duplication).

4.2.3 Prohibited practices and design constraints

Even when the device is not high-risk, certain AI practices are constrained. The AI Act contains prohibitions around specific biometric categorisation inferences and restricts particular deployments of real-time remote biometric identification in publicly accessible spaces for law-enforcement purposes. While a consumer assistive device is not a law-enforcement tool, this still signals that "face recognition in public" and biometric categorisation features should be approached conservatively, with clear technical safeguards and careful scoping to avoid prohibited categorisation inferences and to reduce GDPR special-category exposure.

4.3 GDPR implications for visual data in accessibility applications

4.3.1 Why GDPR almost certainly applies

The device's core operation, capturing video of a user's surroundings, predictably collects personal data when bystanders are visible or identifiable, and also may capture location data and contextual signals. GDPR therefore applies to the provider's processing whenever the provider determines purposes and means (for example, when data is uploaded to servers for inference, improvement, logging, or debugging). A key sensitivity is whether the system processes biometric data "for the purpose of uniquely identifying" a person, which is generally prohibited unless a specific exception applies. Importantly, photographs and video are not automatically "biometric data" merely because they contain faces; they become biometric in the GDPR sense when processed through specific technical means enabling unique identification or authentication. This distinction is central to feature design: "describing a scene" can often be achieved without identifying named individuals.

4.3.2 Controller obligations most salient to an assistive vision product

First, GDPR requires privacy by design and by default: appropriate technical and organisational measures must implement core principles such as data minimisation and ensure that, by default, only data necessary for each purpose are processed and not made accessible to an indefinite number of people. For an assistive camera product, this typically translates into architectural preferences such as on-device inference where feasible, aggressive retention limits, strong access controls, and disabling any "upload by default" flows unless strictly necessary.

Second, security of processing must be appropriate to risk, including measures such as encryption and ongoing testing of safeguards. This requirement is not merely a cybersecurity checkbox: because the device may capture scenes in homes, workplaces, and public spaces, a breach can expose highly sensitive contextual information even if the data are not "special category."

Third, the GDPR's DPIA obligation is highly likely to be triggered. Article 35 requires a data-protection impact assessment when new technologies create likely high risks to individuals' rights and freedoms, with an explicit example being systematic monitoring of a publicly accessible area on a large scale. A wearable camera used in public spaces

is close to the risk profile contemplated by the DPIA regime; accordingly, a DPIA should be treated as a baseline deliverable, not an exceptional exercise.

4.4 MDR relevance: non-medical intention versus disability-compensation reality

The MDR question is less about the internal belief that the product is "not medical" and more about whether the product's intended purpose aligns with the MDR's definition (which includes devices intended, among other things, for compensation for an injury or disability). An assistive device marketed explicitly to compensate for visual impairment can therefore fall within MDR scope even if it does not treat disease. This is the principal regulatory ambiguity for visually-impaired assistive AI in Europe: the disability-compensation claim is often the very value proposition, yet that value proposition can be the trigger for medical-device qualification.

If the product is qualified as an MDR medical device, classification then determines whether third-party conformity assessment is required (and, by extension, whether the AI Act "high-risk" trigger under Article 6(1) is met). Many "assistive information" functions will not resemble diagnostic decision support, but classification depends on the specific claims, the severity of potential harm, and software rules; specialist regulatory assessment is typically required before finalising a compliance strategy.

5 Economic Limits of Scale-First Assistive AI

5.1 Population-Scale Demand Versus System-Level Cost Structures

Visual impairment is a population-scale condition shaped by the sociodemographic gradients discussed earlier (ageing, inequality in service access, and geographic disadvantage). The WHO estimates that at least 2.2 billion people live with near or distance vision impairment worldwide, and at least 1 billion cases are preventable or remain unaddressed (e.g., uncorrected refractive error, untreated cataract). This scale implies that meaningful gains in autonomy require solutions that can expand far beyond small cohorts.

By contrast, many existing assistive pathways carry high per-user costs and recurring system costs (specialist time, training, maintenance, replacement cycles). For example, guide dog organizations commonly report training costs in the tens of thousands of dollars per dog, and refreshable braille displays are often priced in the thousands to over ten thousand dollars. Such figures that reflect real production and service requirements, inherently limit scalability. Even when unit prices fall, total expenditure grows approximately with cost per user × number of users, while public budgets, charitable capacity, and specialist labor do not scale at the same rate. This mirrors the broader assistive-technology landscape: WHO–UNICEF estimate over 2.5 billion people need at least one assistive product, yet nearly 1 billion lack access.

The result is a structural mismatch: even optimistic adoption scenarios for high-cost individualized systems cannot close a gap measured in billions. Addressing population-scale needs requires cost structures and delivery models designed for scale, lower marginal cost per additional user, minimal dependence on scarce specialist time, and compatibility with widely available infrastructure, particularly in the settings where unmet need is greatest.

5.2 Compute-Intensive Architectures and Marginal Cost Growth

Current assistive-vision systems on the market fail to scale primarily due to their cost structure, which is dominated by marginal rather than fixed costs. Each additional user introduces new expenses through hardware, connectivity, servicing, and ongoing inference spend (the latter particularly in cloud-based approaches).

These marginal costs are tightly linked to compute-intensive system designs, and three structural factors contribute most directly to their growth:

-Specialized assistive hardware fails to benefit from economies of scale. Dedicated wearables and custom components are produced in relatively small volumes, limiting cost amortization and keeping unit prices high. Even when devices support offline operation, retail prices remain elevated due to small-batch manufacturing, supply-chain complexity, and ongoing support requirements. As a result, prices do not collapse with demand in the way they do for mass-market consumer electronics.

-Cloud inference transforms assistance into a subscription-based service. Cloud-based scene interpretation shifts costs from a one-time device purchase to recurring, usage-based expenses, including inference, bandwidth, and platform fees. What is nominally an assistive aid thus becomes a metered service: sustained use grows increasingly expensive over time, access becomes contingent on reliable connectivity, and core functionality can degrade or fail entirely in low-connectivity conditions. This pricing model benefits providers but introduces long-term fragility for users.

-Continuous upgrade pressure increases long-run system costs. As AI models evolve rapidly, cloud platforms can update frequently, while deployed hardware becomes locked into compatibility, latency, and privacy constraints. Over time, this creates functional obsolescence: costs rise due to ongoing model updates, monitoring, compliance, and integration, even when the user-facing feature set appears largely unchanged.

Taken together, these dynamics imply that high-compute AI scales more effectively in research than in access. As long as assistive systems carry non-negligible marginal costs, adoption remains concentrated among high-income users and well-resourced institutions, leaving the majority of global need structurally unmet.

5.3 Infrastructure Dependence as an Economic Bottleneck

Infrastructure is often treated as a deployment detail, but in the context of assistive vision it becomes a primary economic constraint. Where the need for assistive technologies is greatest, the supporting infrastructure, such as reliable connectivity, modern devices, stable power, and technical support, is frequently weakest. This creates a structural mismatch between social demand and technical feasibility.

Vision impairment is widespread and projected to grow in the coming decades, with estimates indicating a substantial rise in global sight-loss burden by mid-century. At the same time, access gaps for assistive solutions are already a recognized problem across disability support systems worldwide.

For assistive AI in particular, infrastructure dependence introduces significant and recurring costs through several mechanisms. If inference depends on the cloud, system performance becomes conditional on network availability and latency. In assistive contexts, this does not merely reduce quality of service, it creates instead unpredictable failure modes during mobility, transit, and everyday tasks, where reliability is critical.

Maintenance and energy requirements further increase effective cost. Wearable or always-on camera systems require frequent charging, replacement parts, and technical

servicing. Each requirement introduces hidden expenses and raises abandonment risk, especially in regions where repair ecosystems and technical support are limited. In addition, long-term user success depends on onboarding, accessibility configuration, and reliable update paths, which are processes that demand sustained support rather than one-time setup. Infrastructure-light environments amplify these burdens.

Economically, infrastructure dependence is best understood as cost volatility. Even when a system appears affordable at purchase, the real cost of keeping it functional over time fluctuates with network reliability, device aging, and maintenance capacity. This volatility disproportionately affects the very populations for whom assistive vision technologies would have the greatest potential impact.

5.4 Limits of Proprietary, Closed-System Approaches

Closed ecosystems can deliver tailored experiences, but they impose structural limits on affordability and adaptation that persist even if components become cheaper.

When models, hardware, and interfaces are tightly coupled, organizations cannot swap components to match local constraints (older phones, different languages, limited charging, or different indoor environments). This blocks the most important scaling mechanism for assistive tech, which is local tailoring to the user's unique needs. In addition to this, assistive systems operate in private and public spaces, so if the system's performance limits, failure modes, and data handling practices are not inspectable, it becomes harder for public institutions, NGOs, and researchers to validate safety and privacy claims.

Furthermore, recurring fees can be sustainable for a provider, but they stratify access. In contexts where need is large and purchasing power is low, subscriptions systematically reduce long-term adoption. Even if the hardware price declines over time, cloud dependence, licensing, and proprietary servicing keep the total cost of ownership high.

In contrast, an open, modular approach is not primarily an ethical preference, but rather an economic strategy: it is the most direct path to cost reduction, local repairability, and deployment independence.

5.5 Open and Constraint-First Systems as an Economically Stable Alternative

Open, modular, and constraint-first architectures change the economics of assistive AI because they shift the dominant cost drivers from *marginal costs per user* toward *fixed costs amortized across users*. In scale-first cloud ecosystems, each additional user tends to increase ongoing expenditures (cloud inference, bandwidth, subscription support, proprietary servicing, and device upgrade pressure), making total cost grow roughly with adoption. This scaling law is misaligned with population-level need, especially where purchasing power and infrastructure are limited.

An open and modular system—built to run offline on commodity smartphones—moves the marginal cost of serving an additional user toward near-zero once the software is developed and distributed. Modularity enables local adaptation without renegotiating platform constraints: communities can swap OCR engines, retrain currency classifiers for local denominations, adjust object vocabularies, and tune thresholds without rebuilding the entire stack. In assistive contexts, this adaptability is not a luxury; it is the mechanism by which systems remain usable across languages, scripts, and built environments.

Crucially, "open" here is not framed as ideology. It is an economic stability argument: transparency enables external validation of privacy and safety claims, reuse reduces duplicated engineering across NGOs and public agencies, and decoupling from pro-

prietary infrastructure reduces the risk that core functionality disappears when business models change. When assistance is an accessibility lifeline, long-run continuity and repairability are themselves safety and equity properties.

5.6 Economic Implications for Population-Scale Deployment

Population-scale deployment follows a feasibility condition: solutions must have (i) low marginal cost, (ii) minimal dependence on scarce infrastructure, and (iii) composability that supports localization. Assistive vision demand is measured in the hundreds of millions to billions; systems whose unit economics require expensive wearables, continuous cloud inference, or specialist configuration will remain confined to a thin slice of high-income users and well-resourced institutions.

Constraint-first design is therefore not just an engineering preference but an economic necessity. Offline-first inference reduces recurring connectivity and compute costs; small efficient models reduce battery drain and hardware requirements; and modularity enables incremental improvements without forcing full device replacement cycles. Together, these properties define a cost structure that can plausibly scale: large fixed costs (development, evaluation, documentation) with very low per-user costs and fewer external dependencies.

The practical implication is that the pathway to global impact is not "build the most capable model and later make it cheaper," but "build within the economic and infrastructural envelope from the outset." In assistive vision, the constraint regime is the market: if a system cannot run safely and predictably on widely available devices in low-infrastructure settings, it cannot close the access gap identified earlier.

6 Synthesis and Design Direction

6.1 What have we got so far?

Over the course of the previous sections, we moved from a broad motivation to a concrete technical plan. This subsection documents the technical work already completed at the design level: the constraints we formalized, the modules we committed to, and the reliability behaviors we treat as non-negotiable. We have already defined the system's operating assumptions, functional scope, and failure policy in a way that makes subsequent implementation and evaluation a matter of execution rather than conceptual uncertainty.

6.1.1 From constraints to an explicit system specification

Based on the deployment realities established earlier (noise, motion blur, intermittent connectivity, limited compute/energy, high cost of failure), we formalized a constraint-first specification for the prototype. The system is designed to be: -Offline-first: core functions remain available without cloud inference. -Low-resource compatible: bounded assumptions on compute availability, memory capacity, and end-to-end latency are treated as first-order design constraints rather than post-hoc deployment issues. -Reliability-oriented: prioritizes predictable behavior and graceful degradation over maximal benchmark accuracy, reflecting the safety-critical nature of assistive feedback. -With low cognitive load: interaction is minimized and information output is throttled to reduce overload, consistent with findings that assistive systems can fail when perception outputs are not turned into actionable guidance.

The pipeline must remain usable under degraded camera inputs (lighting variance, blur, occlusion) and under limited compute, rather than being optimized only for ideal conditions.

6.1.2 Modular assistive pipeline

We adopted a standard assistive architecture that goes from sensing, to perception, to feedback as the organizing structure for implementation, because it maps directly to what the literature identifies as common and practical in deployed assistive systems.

1. Sensing (input): smartphone camera (primary), with optional use of IMU signals for stabilization and motion cues.
2. Perception (modules): OCR, object detection, scene recognition, currency recognition, and motion/obstacle cues.
3. Feedback (output): audio by default, with optional haptics; messaging is event-based and mode-specific (not continuous narration), aligning with evidence that feedback overload and latency reduce usability.

This modularity avoids a monolithic architecture and keeps each task independently optimizable and replaceable.

6.1.3 Reliability policy

Outputs are confidence-aware: the system can refuse to answer if confidence is low (especially for currency), and instead issues recovery prompts. Moreover, the system is designed to degrade safely, reducing scopre under poor conditions (e.g., reading the clearest/largest text line; identifying fewer but more confident objects), rather than richer but inaccurate descriptions.

6.2 Product Concept and Intended User Experience

The future product is planned as an offline assistive vision application, initially deployed on a smartphone (with an option to extend to wearable hardware later on). The proposed product is designed to support visually impaired users in common everyday tasks by providing short audio feedback.

The core idea is to minimise cognitive load and interaction complexity:

- A small number of task modes, each optimised for a specific use case (e.g., reading text or finding objects).
- One-touch interaction to trigger perception tasks.
- Event-based feedback, rather than a constant audio message (only provide feedback when something important changes or is requested by the user).
- Confidence-aware output, where the system signals uncertainty and guides the user to provide a clearer camera view, rather than providing a guessed answer.

6.3 Interface and Interaction Design

The product UI is envisioned as a simple, accessibility-first interface with four core models accessible via large, high-contrast buttons.

1. **Read Mode (Optical Character Recognition - OCR)**
   This mode supports the perception of printed text in everyday contexts (e.g., signs, labels, menus, documents). When the user points the camera at the target text and specifically selects this mode, the system recognises and reads out the most

noticeable textual element - such as the largest or clearest line - with the option to require more text output if necessary.

2. **Find Mode (Object Identification)**
   The purpose of this mode is to identify common, established object categories that are important for interaction and navigation. The system recognises target objects upon activation and uses coarse positional descriptors (e.g., left/center/right and near/far) to offer brief audio messages explaining their presence and approximate spatial placement.

3. **Money Mode (Currency Recognition)**
   The focus of this mode is to identify banknotes and their denominations. The user initiates the mode and displays a steady image of the currency. Only when the confidence levels are satisfied does the system provide a denomination label. When there is not enough certainty, the system does not provide a definitive output and instead asks the user to clarify the input.

4. **Move Mode (Motion and Obstacle Cues)**
   This mode supports lightweight mobility in offline and low-resource scenarios. Once activated, the system sends out cautious notifications using brief audio messages and optional haptic feedback while keeping an eye out for motion or nearby obstacles.

Across all interaction modes, the interface includes the following global controls and feedback mechanisms:

- Repeat the last output, allowing users to re-access recent feedback without re-running the interface.
- Mute or quiet mode, allowing users to temporarily shut off audio output.
- Guidance prompts, such as "move closer", "increase lighting," or "hold steady," are used to ensure the input quality is sufficient.

6.4 Model Selection

The current technical research supports a mixed approach: small deep models where necessary and classical computer vision where reliability and computational efficiency dominate.

6.4.1 Optical Character Recognition (OCR)

Tesseract OCR with OpenCV-based preprocessing is selected because of its ability to run solely on CPU, perform well on huge printed text, and work with offline-first deployment scenarios. To increase robustness under changing input conditions, preprocessing techniques such as grayscale conversion, thresholding, and geometric normalisation are applied. A more advanced approach would be to use detection-recognition OCR (e.g., OpenOCR or PaddleOCR-style architectures). These models are considered for future improvements targeting more complex scene text.

6.4.2 Object Detection

We plan to use lightweight object detectors, including MobileNet-SSD and compact YOLO variants, which are limited to a predetermined selection of assistive-relevant object classes. In order to facilitate effective on-device inference, deployment is anticipated through OpenCV DNN, ONNX Runtime, or TensorFlow Lite. These models offer real-time or nearly real-time performance with explicit control over inference frequency, memory footprint, and latency.

6.4.3 Scene Recognition

A two-tiered approach is proposed for scene-level interpretation. For consistent, coarse-grained scene classification (e.g., indoor/outdoor, street, kitchen), Places365-trained convolutional neural networks (CNNs) provide a low-cost baseline. Despite their significantly greater computational requirements, CLIP or OpenCLIP- based model are regarded as optional semantic enhancers in user-invoked modes (such as environment description).

6.4.4 Currency Recognition

We plan to use a tiny, fine-tuned convolutional classifier (MobileNetV2 or MobileNetV3), which is trained on currency images specific to a given region. Currency recognition is considered a high-stakes, specialised categorisation task. Strict confidence criteria are applied to model outputs, and in situations of doubt, absences and recapture reminders are used to reduce the possibility of inaccurate output.

6.4.5 Motion and Obstacle Cues

Classical computer vision approaches, which provide predictable behaviour and low processing overheads, are the foundation of the system's motion and obstacle recognition. Under resource-constrained circumstances, relative motion, scene changes, and coarse structural signals are detected via optical flow (Lucas-Kanade), background subtraction, and edge detection. Due to their greater resource requirements, more computationally demanding methods - like RAFT-based optical flow and MiDaS monocular depth estimation - are regarded as optional research expansions and are assessed experimentally rather than being incorporated into the default workflow.

Bibliography

1. OrCam Techologies. (2024). OrCam MyEye 3 Pro - Revolutionize Your Vision with Cutting-Edge AI Technology. [online] Available at: https://www.orcam.com/en-us/orcam-myeye-3-pro

2. Be My Eyes. (2025). Be My Eyes Smartglasses. [online] Available at: https://www.bemyeyes.com/be-my-eyes-smartglasses/

3. Letsenvision.com. (2024). Envision Glasses. [online] Available at: https://www.letsenvision.com/glasses/home

4. eSight Eyewear. (2025). eSight Go | eSight Eyewear. [online] Available at: https://www.esighteyewear.com/esight-go/

5. European Data Protection Board. (2019). Guidelines 3/2019 on processing of personal data through video devices.

6. European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L 119.

7. European Parliament and Council of the European Union. (2017). Regulation (EU) 2017/745 of 5 April 2017 on medical devices. Official Journal of the European Union, L 117.

8. European Parliament and Council of the European Union. (2024). Regulation (EU)

2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L (2024/1689).

9. European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation).

10. European Union. (2017). Regulation (EU) 2017/745 on medical devices (MDR).

11. European Union. (2019). Directive (EU) 2019/882 (European Accessibility Act) (summary).

12. ETSI. (2025). EN 301 549 V4.1.0: Accessibility requirements for ICT products and services.

13. International Agency for the Prevention of Blindness (IAPB). (2025). The Value of Vision (Full Report).

14. Lancet Global Health Commission on Global Eye Health. (2021). Commission Report (PDF).

15. Smith-Kettlewell Eye Research Institute. (n.d.). Talking Signs project overview.

16. World Health Organization. (2019). World report on vision.

17. World Health Organization. (2023). Blindness and visual impairment (Fact sheet).

18. WHO & UNICEF. (2022). Global report on assistive technology (UNICEF landing page; JAMA summary).

19. Kurzweil Technologies / American Foundation for the Blind. (n.d.). Kurzweil Reading Machine background (assistive reading technology).

20. Real, S., et al. (2019). Navigation Systems for the Blind and Visually Impaired: Past Work and Future Trends. Sensors.

21. World Health Organization (2023), Blindness and visual impairment (Fact sheet).

22. WHO–UNICEF (2022), Global report on assistive technology.

23. Bourne R, et al. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. Lancet Glob Health. 2020. Accessed via the IAPB Vision Atlas: visionatlas.iapb.org.

24. Guide Dogs of America, FAQ: training cost estimates.

25. American Foundation for the Blind, Refreshable braille display cost ranges.