The Escalator Problem: Identifying Implicit Motion Blindness in AI for Accessibility

Xiantao Zhang

Beihang University

zhangxiantao@buaa.edu.cn

Abstract

*Multimodal Large Language Models (MLLMs) hold immense promise as assistive technologies for the blind and visually impaired (BVI) community. However, we identify a critical failure mode that undermines their trustworthiness in real-world applications. We introduce the **Escalator Problem**—the inability of state-of-the-art models to perceive an escalator's direction of travel—as a canonical example of a deeper limitation we term **Implicit Motion Blindness**. This blindness stems from the dominant frame-sampling paradigm in video understanding, which, by treating videos as discrete sequences of static images, fundamentally struggles to perceive continuous, low-signal motion. As a position paper, our contribution is not a new model but rather to: (I) formally articulate this blind spot, (II) analyze its implications for user trust, and (III) issue a call to action. We advocate for a paradigm shift from purely semantic recognition towards robust physical perception and urge the development of new, human-centered benchmarks that prioritize safety, reliability, and the genuine needs of users in dynamic environments.*

1. Introduction

Vision foundation models (VFMs) and Multimodal Large Language Models (MLLMs) have ushered in an era of unprecedented progress in artificial intelligence, with the potential to revolutionize assistive technologies. For the blind and visually impaired (BVI) community, these advancements promise a future where AI can act as a "visual interpreter," describing complex scenes, identifying objects, and empowering greater independence in daily life. This vision aligns perfectly with the core mission of the accessibility research community: to harness cutting-edge technology to dismantle barriers and foster a more inclusive world. The prospect of MLLMs acting as real-time, conversational "visual interpreters" via live video is no longer science fiction but a rapidly commercializing reality, as exemplified by
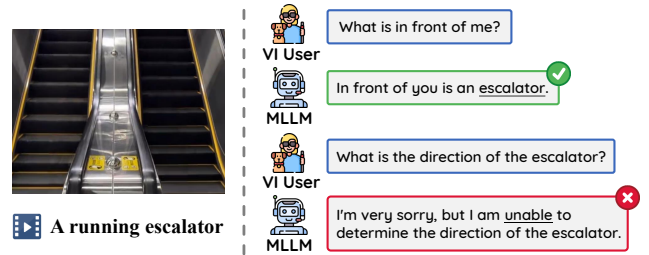

Figure 1. The *Escalator Problem*: A Real-World Failure of Motion Perception. An interaction with the MLLM demonstrates a critical failure. While the model correctly identifies the object as an escalator, it is unable to determine its direction of motion, a vital piece of information for a BVI user's safe navigation. This highlights the concept of *Implicit Motion Blindness*.

models like GPT-4o [14] and Qwen2.5-Omni [23]. These systems, capable of fluid dialogue about a user's live surroundings, represent an exciting frontier.

However, despite this encouraging progress, a critical gap persists between current model capabilities and the demands of real-world navigation. In this paper, we identify a fundamental failure mode using the *Escalator Problem* as a canonical, illustrative case: determining the direction of a moving escalator from a first-person video. To demonstrate this blind spot, we presented video footage of a standard escalator to leading MLLMs. Their responses were consistently alarming; as illustrated in Figure 1, even top-tier models often fail this seemingly trivial task. They tend to default to describing a static scene or explicitly state their inability to perceive motion. This simple observation forms the crux of our investigation.

This paper posits that this failure is not an isolated glitch but rather a symptom of a deeper, systemic issue we term *Implicit Motion Blindness*. We argue that the dominant frame-sampling paradigm in video understanding—which processes videos as a sequence of discrete, static images—is ill-equipped to perceive low-signal, continuous motion. The crucial directional information in the

escalator video does not reside within any single frame, but exists implicitly between them, in the subtle, continuous flow that is effortlessly perceived by the human visual system. The models' inability to capture this flow reveals a critical blind spot for tasks vital for safe and independent navigation.

As a position paper, our contribution is not a new model or dataset but rather the identification and articulation of this critical blind spot to steer the community's focus. We make the following three contributions:

I We identify and formally articulate the *Escalator Problem* as a canonical, real-world failure case that highlights the limitations of current MLLMs in perceiving continuous, low-signal motion.

II We analyze the broader implications of this *Implicit Motion Blindness* for the BVI community, arguing that it represents a significant barrier to building trustworthy assistive technologies and can erode user trust in real-world deployments.

III We issue a call to action for the research community to shift its focus from purely semantic recognition towards robust physical perception, and call for the co-development of new, human-centric evaluation paradigms that prioritize safety, reliability, and trust in assistive AI.

2. Background and Related Work

2.1. Video Perception in MLLMs

The primary method for MLLMs to process video is through **sparse frame sampling**. This technique involves selecting a small, discrete subset of frames to represent the entire video sequence, drastically reducing the number of visual tokens for the model to process. Common methods include uniform sampling and frames-per-second (FPS) sampling [1, 10, 20–22].

While computationally efficient, this approach is fundamentally lossy. By discarding most frames, sparse sampling severs the continuous dynamic information that constitutes motion and temporal progression. Consequently, models can miss crucial events and struggle with tasks requiring fine-grained spatiotemporal understanding [16]. For a typical 5-minute video, a model limited to 64 frames would sample at a rate of one frame every five seconds, making detailed motion analysis nearly impossible [8]. This method treats a video as an unordered "bag of frames," sacrificing temporal integrity for manageable data size.

2.2. Video Understanding Benchmarks

Leading video understanding benchmarks, such as Kinetics [9] and ActivityNet [6], have been pivotal in advancing research. These datasets primarily focus on **explicit action recognition**, where models are tasked with classifying activities like "running" or "playing guitar."

However, these benchmarks often have a "*static appearance bias*" [15]. Many actions can be identified from a single frame, such as classifying diving by seeing a pool, often making temporal information redundant [24]. This is confirmed by experiments where models perform well even with shuffled frames, showing that true temporal understanding is not always being evaluated [5]. Consequently, these benchmarks test for high-level action classification where visual clues are often in a few keyframes, rather than assessing a model's ability to reason about events over time.

2.3. AI for Accessibility

AI has significantly advanced assistive technology for the visually impaired, with a global population of over 2.2 billion [13]. Applications like Microsoft's Seeing AI [12] and Google Lookout [7] excel at static perception tasks, including optical character recognition (OCR), product identification, and scene description. Wearable devices such as Envision Glasses [4] and the NOA vest [2] offer more continuous navigation support by providing obstacle detection and GPS navigation.

Despite these advances, a critical gap remains in handling dynamic, unpredictable environments [17]. Most assistive technologies focus on obstacle detection and static scene description, falling short of the nuanced understanding required for safe mobility [13]. They can identify an "escalator" but struggle to interpret its dynamic context, such as whether it is broken and must be walked on. This disconnect between low-level perception and high-level situational awareness is a major barrier to true navigational independence and safety.

3. Why Models Can't See What's Moving

The failure of sophisticated MLLMs on the Escalator Test is not a matter of insufficient parameters or inadequate training data in the conventional sense. Rather, it exposes a fundamental misalignment between the strategy models use to "see" and the nature of visual information in the physical world. To understand the failure, we must first appreciate the profound difference between how humans and current AI models perceive continuous motion.

3.1. The Human Advantage: Perceiving the Flow

For a human observer, determining the escalator's direction is a pre-attentive, almost instantaneous process. We do not analyze the scene as a series of static photographs. Instead, our visual system is exquisitely tuned to perceive *optical flow*—the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and the scene. We perceive a continuous stream of information. Our brain effortlessly detects the
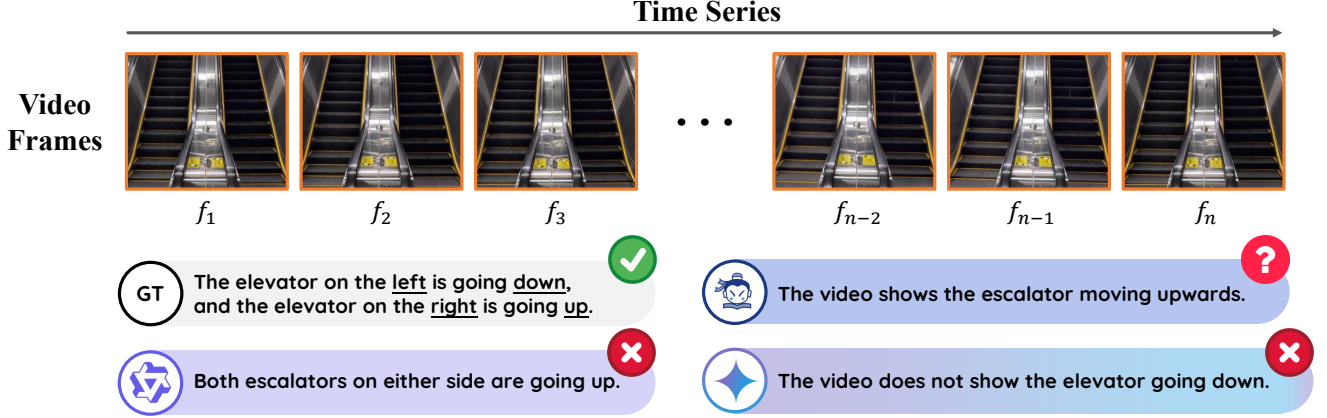
Figure 2. An Illustrative Example of *Implicit Motion Blindness*. When presented with a video of two adjacent escalators, current MLLMs fail to correctly describe the motion. The Ground Truth (GT) indicates opposite directions, but model outputs are incorrect, either guessing the wrong direction or denying the presence of motion altogether. This empirically validates the limitations of the current frame-sampling paradigm for tasks requiring fine-grained temporal understanding.

subtle relative movements: the handrail gliding smoothly past a stationary pillar, the seamless cascade of the steps, which creates a waterfall effect, and the gentle parallax shift of background elements relative to the foreground.

This perception is holistic and relational. The directional signal is not contained in the metallic texture of a single step but in the collective, coherent movement of all steps relative to the static surroundings. It is this perception of flow, not the recognition of objects, that allows for immediate and confident judgment. This holistic, physics-aware process stands in stark contrast to the frame-by-frame object identification paradigm of MLLMs. The visual evidence is clear and unambiguous precisely because our brains are built to process temporal dynamics as a primary source of information about the world.

3.2. The Model's Blind Spot: Frame-Level Myopia

In stark contrast, an MLLM processing the same video operates with a form of **frame-level myopia**. The dominant video understanding paradigm involves sampling discrete frames from the video at a certain rate and feeding these static images into the model for analysis. As illustrated in Figure 3, when viewed in isolation, these frames are nearly indistinguishable. The subtle displacement of the escalator steps from one frame to the next is often smaller than the visual noise or may be entirely lost if the sampling rate is too low.

To illustrate this limitation, consider the stark contrast between this Escalator Problem and a high-signal motion event, such as a person walking across a room. In the latter case, even with sparse frame sampling, the significant change in the subject's silhouette, position, and the occlusion of background elements between frames provides a

strong, unambiguous motion signal. The model can easily detect this change. The escalator, however, represents a near-worst-case scenario for this paradigm: its motion is characterized by uniform, continuous, and texturally repetitive flow. The visual evidence of motion is low-signal and distributed evenly across the entire object, rather than being concentrated in a distinct, moving entity. It is this specific failure to perceive motion in the absence of a high-signal, discrete moving object that defines the core of Implicit Motion Blindness.

The model, therefore, is presented with a series of snapshots that essentially say: "This is an escalator," "This is the same escalator," "This is still the same escalator." Its training on vast datasets like ImageNet [3], COCO [11], or even large-scale video datasets has optimized it for semantic recognition—identifying objects, scenes, and discrete actions. However, it has not been sufficiently trained to understand the fine-grained physical processes that generate the visual data. The model excels at answering "What is this?" but falters when asked, "How is it behaving?". The critical directional signal, which is encoded in the temporal relationship between frames, is lost in the sampling process. This architectural blind spot is the root cause of its failure on the Escalator Test, as empirically demonstrated in Figure 2, where state-of-the-art models fail to correctly describe the motion of two adjacent escalators.

3.3. The Frame-Sampling Paradigm

To deconstruct the "*frame-level myopia*" with technical precision, we can formalize the general paradigm through which modern MLLMs process video. The core objective of such a model is to learn a conditional probability distribution $P(Y|V,T)$, which is the probability of generating a

[Scene] Sunlight bathes the grass as a puppy runs joyfully, radiating life and happiness.
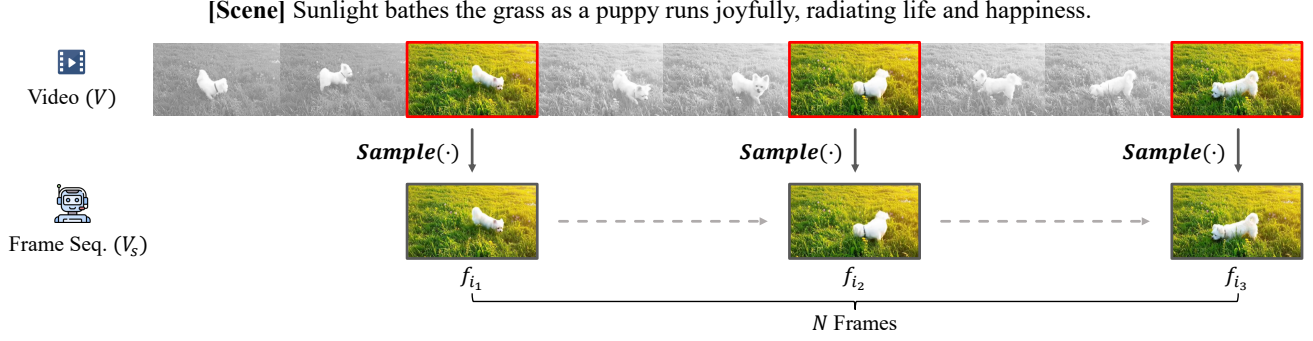

Figure 3. The Frame-Sampling Paradigm as the Source of *Motion Blindness*. This diagram illustrates how MLLMs process video by sampling a sparse sequence of frames (V_s) from the original video (V). For high-level semantic descriptions, this is often sufficient. However, for continuous, low-signal motion like an escalator, the subtle displacement between any two sampled frames ($f_{i_1}, f_{i_2}, f_{i_3}$) is often lost, leading the model to perceive a series of static, nearly identical images rather than a continuous flow.

target text sequence Y given an input video V and a text prompt T. This process involves four stages.

Symbol Definitions.

V The raw input video, represented as a tensor $V \in \mathbb{R}^{T_v \times H \times W \times C}$, where T_v is the total number of frames, H, W are the frame dimensions, and C is the number of color channels.

T The input text prompt, a sequence of M tokens $T = (w_1, w_2, \ldots, w_M)$.

Y The generated output text, a sequence of L tokens $Y = (y_1, y_2, \ldots, y_L)$.

d The model's internal feature dimensionality.

Stage 1: Video Representation. This initial stage converts the continuous video signal into a discrete, ordered sequence of image patches. This is where the critical information loss occurs.

Frame Sampling. A sampling function, Sample$(\cdot)$, extracts N frames from the full video V to form a frame sequence V_s.

$$V_s = \text{Sample}(V, N) = (f_1, f_2, \ldots, f_N) \quad (1)$$

where each frame $f_i \in \mathbb{R}^{H \times W \times C}$. Crucially, $N \ll T_v$.

Patchification. A patching function, Patch$(\cdot)$, divides each frame f_i into K smaller image patches.

$$\text{Patch}(f_i) = (p_{i,1}, p_{i,2}, \ldots, p_{i,K}) \quad (2)$$

where each patch $p_{i,j} \in \mathbb{R}^{P \times P \times C}$, and P is the patch size.

Stage 2: Feature Encoding. This stage maps the discrete visual and textual data into a unified, high-dimensional continuous vector space.

Visual Encoding. A visual encoder E_v (e.g., a Vision Transformer, ViT) transforms the sequence of all patches from all sampled frames into a sequence of d-dimensional visual features Z_v.

$$Z_v = E_v(V_s) \in \mathbb{R}^{(N \times K) \times d} \quad (3)$$

Text Encoding. A language encoder E_l (e.g., a Transformer Encoder) converts the prompt T into a sequence of d-dimensional text features Z_l.

$$Z_l = E_l(T) \in \mathbb{R}^{M \times d} \quad (4)$$

Stage 3: Multimodal Fusion. This core stage aligns and integrates information from the different modalities to form a unified, cross-modal understanding.

We define a generic fusion module F_m that takes the independent visual and text feature sequences, Z_v and Z_l, as input and outputs a fused hidden state sequence, H_{fused}.

$$H_{\text{fused}} = F_m(Z_v, Z_l) \quad (5)$$

Functionally, F_m "anchors" textual concepts to visual evidence within a shared semantic space by modeling complex intra-modal and inter-modal relationships. This can be implemented via co-attention mechanisms, dedicated fusion layers, or by concatenating and passing features through a deep Transformer. For our purposes, we abstract away the specific implementation. The final output, $H_{\text{fused}} \in \mathbb{R}^{S \times d}$, is a context-aware representation ready for downstream tasks, where S is the length of the fused sequence.

Stage 4: Autoregressive Generation. Based on the fused representation H_{fused}, an autoregressive decoder generates the final output text Y token by token. Following the chain rule of probability, the generation process is formalized as:

$$P(Y|V, T) = \prod_{t=1}^{L} P(y_t|y_{<t}, H_{\text{fused}}) \quad (6)$$

At each timestep t, the probability distribution for the next token is determined by a function G (typically a linear layer followed a Softmax function):

$$P(y_t|y_{<t}, H_{\text{fused}}) = G(H_{\text{fused}}, y_{<t}) \tag{7}$$

The model samples a token y_t from this distribution, appends it to the input for the next step, and repeats the process until an end-of-sequence token is generated or a maximum length is reached.

This formalization makes the root of the Implicit Motion Blindness explicit: the irreversible loss of temporal continuity occurs at the very first step, in the Sample(V, N) operation. No matter how sophisticated the subsequent fusion and generation stages are, they cannot recover the fine-grained motion information that was discarded before they ever saw the data.

4. The Ripple Effect: Beyond Escalators, A Crisis of Trust

The inability of advanced AI to pass the *Escalator Test* would be an academic curiosity if it were an isolated phenomenon. However, its true significance lies in what it represents: a crack in the foundation of our approach to building assistive technologies for the real world. This Implicit Motion Blindness is a systemic vulnerability that manifests across a wide spectrum of everyday scenarios. The ultimate consequence of this vulnerability is not just functional failure, but the erosion of the single most critical element for any assistive tool: **user trust**.

4.1. A Spectrum of Implicit Motion Challenges

The Escalator Problem is a stand-in for a broader class of environmental interactions where perceiving subtle, continuous motion is not merely helpful, but essential for safety and social integration. In each case, the core challenge remains the same: the critical information is encoded in the dynamic flow of the environment, not in the static identity of its components. Consider the scenarios outlined in Table 1. Navigating a public space like an airport or a busy street requires constant, implicit interpretation of motion cues that current MLLMs are ill-equipped to handle.

For a BVI user, an AI assistant that cannot distinguish the direction of a baggage carousel, the subtle sway of a revolving door, or the dominant flow of a crowd is not just inconvenient; it is functionally deficient. It fails to provide the very layer of situational awareness that such a tool is meant to deliver. The model might correctly identify a crowd of people, but it fails to provide the actionable insight: "the crowd is moving towards your left." It might see a revolving door, but cannot answer the crucial question: "is it safe to enter now?". This gap between semantic description and

perceptual understanding transforms the AI from a potential navigator into a mere commentator, narrating a scene without providing the means to act within it.

4.2. The Consequence: A Crisis of Trust

This leads to the most profound implication of our findings: a crisis of trust. An assistive tool that is predictably unreliable in common, everyday situations is worse than no tool at all. Trust in an assistive device is not built on its exceptional performance on benchmark tasks; it is forged through consistent, reliable behavior in the mundane, messy reality of daily life. When a user discovers their AI assistant cannot solve a problem as fundamental as the *Escalator Test*, their confidence in the system's ability to handle more complex or higher-stakes situations—like identifying a safe gap in traffic to cross a street—is fundamentally undermined.

This issue speaks directly to the core challenges of **human-centered evaluation** and the **real-world deployment of AI-powered assistive systems**. A system that is 99% accurate on a dataset of discrete object labels but fails on 100% of implicit motion tasks is, from a human-centered perspective, a brittle and therefore untrustworthy system.

This brittleness creates more than just functional unreliability; it imposes a significant **cognitive load** [18] on the BVI user. In human-computer interaction, a primary goal of assistive technology is to reduce this load. However, a predictably unpredictable system forces the user into a constant state of vigilance, compelling them to second-guess the AI's output and determine which perceptions are reliable. This directly contradicts the technology's core purpose. This transforms the interaction from a seamless assistive experience into a stressful cognitive puzzle, where the user must constantly deduce: "Is this an object the AI is good at, or a motion it's blind to?"

Furthermore, we must distinguish between trust, a user's general belief in the system's competence and integrity, and reliance, the decision to depend on the system in a specific context. The Escalator Problem erodes fundamental trust at a systemic level. Consequently, even if the model performs perfectly on 99% of static object recognition tasks, a rational user cannot rely on it for any task involving dynamic situational awareness. This transforms the AI from a potential partner into a mere "situational commentator," one whose observations require constant, stressful verification, and whose failures risk not only physical safety but also social awkwardness and a diminished sense of independence.

Therefore, Implicit Motion Blindness is not simply a technical limitation to be incrementally improved. It is a fundamental barrier to establishing the user trust necessary for adoption and effective use. Without addressing this core perceptual deficit, we risk developing technologies that look impressive in demonstrations but fail silently and dangerously in the hands of the very people we aim to empower.

Scenario	Challenge for AI (*Implicit Motion*)	Potential Consequence for BVI User
Escalator / Travelator	Perceiving direction of linear motion from a first-person perspective.	Wasted time, disorientation, social awkwardness, risk of falling.
Revolving Door	Distinguishing between moving, stationary, or blocked states.	Attempting to enter a static or blocked door, risk of collision.
Crowd Flow	Determining the dominant direction and speed of a moving crowd.	Inability to navigate efficiently, moving against the flow, social friction.
Flowing Water / Puddles	Detecting subtle surface movement indicating flow or disturbance.	Risk of stepping into moving water (e.g., gutter, drain runoff), wet feet, potential slips.
Automatic Sliding Doors	Perceiving the subtle initial movement indicating the doors are opening.	Hesitation, walking into a closed door, awkward waiting.
Baggage Carousel	Identifying direction and speed of the conveyor belt.	Difficulty in positioning oneself to retrieve luggage, increased stress.

Table 1. A Spectrum of Implicit Motion Challenges for Assistive AI. The Escalator Problem is representative of a broader class of real-world scenarios where understanding implicit, continuous motion is critical for safe and effective navigation for BVI users. This table outlines several common situations where current AI approaches are likely to fail, highlighting the systemic nature of the issue.

The path forward requires us to acknowledge this crisis and re-evaluate the foundational principles upon which our assistive systems are built.

5. A Call for a New Direction: From Recognition to Perception

The discovery of *motion blindness* necessitates more than an incremental fix; it demands a fundamental re-evaluation of our research trajectory in AI for accessibility. Continuing to scale models on ever-larger datasets of labeled objects and discrete actions will not solve the Escalator Problem. The solution lies in a paradigm shift: we must guide our models to evolve from engines of **recognition** to systems of **perception**. A system of recognition asks, "What is this object?"; a system of perception asks, "How is this scene behaving, and what does that mean for me?". For an assistive AI, the latter is immeasurably more important.

5.1. Move Beyond Frame-Based Semantics

This paper's central position is that the computer vision community, particularly in its pursuit of assistive technologies, must consciously move beyond the limitations of semantic frame description and prioritize the development of true motion perception. The frame-sampling approach, while powerful for cataloging the contents of a video, is an abstraction that breaks the continuity of reality. It treats time as a slideshow, not a stream. This is a critical flaw when building tools designed to help a user navigate a dynamic, physical world.

We argue that Implicit Motion Blindness is a direct and inevitable consequence of this flawed abstraction. To build systems that are robust, trustworthy, and genuinely useful in real-world settings, we must treat motion not as an afterthought derived from comparing static frames, but as a primary and fundamental source of information. This requires a shift in architectural priorities, data collection philosophies, and the very definition of what it means for a model to understand a video.

5.2. A Call for Human-Centered Benchmarks

The emergence of real-time, conversational MLLMs makes the development of new evaluation paradigms not just an academic exercise, but a pressing safety imperative. The current landscape of benchmarks, while valuable for gauging progress in classification, detection, and semantic question-answering, is insufficient for evaluating real-world assistive performance.

Therefore, we issue a call for a community-wide effort to rethink evaluation itself. Table 2 provides a comparison that highlights the key differences between current benchmarks and our proposed human-centered approach. We advocate for the development of a new class of evaluation protocols, co-designed with the BVI community, HCI experts, and accessibility professionals. These new benchmarks should:

- **Prioritize Real-World Tasks:** Move beyond abstract VQA and incorporate tasks directly relevant to navigation and safety, such as those characterized by Implicit Motion Blindness, including determining escalator direction, crowd flow, and door state.
- **Measure Trustworthiness, Not Just Accuracy:** Develop metrics that capture system reliability and predictability. A model that correctly identifies an escalator's direction 100% of the time is more valuable than a model that can name 10,000 objects but is unreliable on this critical task. This may involve measuring not just the correctness of an answer, but also its confidence, consistency, and its ability to signal uncertainty (i.e., to "know

what it doesn't know").

- **Embrace Egocentric and Continuous Evaluation:** Shift from evaluating performance on curated, third-person video clips to assessing continuous performance in unstructured, egocentric first-person video streams, which more accurately reflect the input of a wearable assistive device.

5.3. Promising Avenues for Exploration

Solving this challenge will require innovation and an openness to exploring architectures and modalities beyond current MLLMs. While we do not claim to have the solution, we encourage researchers to investigate several promising avenues that prioritize the physics of motion over the semantics of objects.

Hybrid Approaches Before the deep learning era, classical computer vision techniques like optical flow were purpose-built to estimate motion fields between frames. Their efficacy for the Escalator Problem stems from a fundamentally different paradigm: instead of recognizing semantic objects, they directly compute pixel-level displacement vectors between frames, a process that is agnostic to what the object *is*. This makes them inherently robust to challenges where MLLMs are confounded by the repetitive, semantically-identical appearance of the steps from one frame to the next. We propose reintegrating these robust and efficient algorithms into modern architectures. Instead of being replaced by end-to-end models, optical flow could serve as a strong motion prior, a parallel processing stream, or a verification mechanism that grounds an MLLM's abstract understanding in the physical reality of pixel movement.

A potential implementation could be a two-stream fusion architecture. One stream would consist of the conventional MLLM, processing sparsely sampled frames to excel at semantic understanding—answering "What is this object?". In parallel, a second, lightweight stream dedicated to motion analysis, perhaps using a modern optical flow network like RAFT [19], would process denser frame pairs to generate a robust motion vector field, answering "How is this scene behaving?". The outputs could be fused, where the motion stream acts as an attentional mechanism, guiding the MLLM to focus on dynamic regions, or as a verifier that fact-checks the MLLM's generated descriptions against physical motion evidence.

New Sensing Modalities The very sensor we use—the conventional frame-based camera—contributes to the problem. We should explore alternative sensors like **event cameras**. Unlike traditional cameras that capture frames at a fixed rate, event cameras are bio-inspired sensors that asynchronously report pixel-level brightness changes. They are

inherently motion-detectors. Their data is sparse, has high temporal resolution, and requires low power consumption, making them nearly ideal sensors for detecting subtle motion in a wearable, real-time assistive device.

Adopting these sensors, however, presents its own fundamental research challenges that the community must address. Event data, a sparse stream of asynchronous events, is incompatible with the dense, grid-based input expected by architectures like Vision Transformers. This necessitates a new line of inquiry: should we develop novel, non-Transformer architectures natively suited to processing such sparse event streams, or should we focus on creating sophisticated methods to "render" or "tokenize" event data into a dense, frame-like representation that existing MLLMs can comprehend without losing the crucial high-temporal-resolution information?

Physics-Informed Learning A more ambitious, long-term direction is to develop models with an innate, intuitive understanding of physics. This involves creating learning frameworks where models are not just trained on visual data, but are constrained or guided by the fundamental laws of motion, gravity, and continuity. A physics-informed model wouldn't need to have seen a million escalators; it would "understand" that a rigid set of stairs in continuous motion must be moving in a coherent direction.

6. Limitations and Future Discourse

In advocating for this paradigm shift, we acknowledge several important considerations and potential counterarguments.

First, one might argue that advanced prompt engineering could mitigate this issue. While a carefully crafted prompt (e.g., "Describe the motion of the steps relative to the stationary handrail") might yield better results in isolated cases, we contend it is not a fundamental solution. This approach shifts the cognitive burden to the BVI user, requiring them to diagnose the AI's failure mode and formulate a precise query mid-navigation, which runs counter to the goal of a seamless assistive experience. More importantly, if the essential motion information is already lost in the initial $\text{Sample}(V, N)$ operation, no amount of clever prompting can fully recover it.

A second counterargument might point to alternative temporal processing pipelines, such as employing denser or adaptive frame sampling. While these methods may offer marginal improvements, they do not fundamentally resolve the Escalator Problem. For one, they often introduce significant computational overhead, a critical barrier for real-time applications on power-constrained wearable devices. More fundamentally, for textures as repetitive and uniform as escalator steps, even denser sampling can fail to provide a clear directional signal to a model optimized for semantic

Feature	Current Benchmarks	Proposed Human-Centered Benchmarks
Primary Task	Semantic Action Recognition	Fine-grained Physical Perception
Motion Type	High-Signal / Discrete Actions	Low-Signal / Continuous Motion
Static Bias	High (Keyframe-based)	Low (Requires Temporal Reasoning)
Evaluation Metric	Classification Accuracy	Trust, Reliability, Safety
Data Source	Curated Third-Person Video	Egocentric First-Person Streams

Table 2. A Comparison Between Current and Proposed Human-Centered Benchmarks. This table contrasts the focus of existing video understanding benchmarks with the proposed new paradigm. The shift moves from high-level action classification in curated videos towards evaluating trustworthiness and safety on continuous, first-person video streams, directly addressing the shortcomings that lead to Implicit Motion Blindness.

recognition. This suggests the issue is less about the *quantity* of temporal data and more about the *quality* and *nature* of its representation. The problem lies not just in sampling, but in a paradigm that prioritizes *what* an object is over *how* it is behaving.

Furthermore, the call to prioritize physical perception should not be misinterpreted as a call to abandon semantic richness. The ideal future system is not one that sacrifices the MLLM's powerful descriptive abilities but one that gracefully integrates physical and semantic understanding. The key challenge for future research will be to resolve the potential trade-off between these two modes of perception, creating a model that both understands the physics of a scene and can describe it with human-like nuance.

Finally, while our position is that this problem is not merely a matter of scale, we acknowledge that the interplay between model size and emergent abilities is complex. Future work should critically examine whether simply scaling models further on undifferentiated data exacerbates this issue by reinforcing static biases, or if specific, physics-aware training regimes at scale can lead to a genuine, generalizable understanding of motion.

7. Conclusion

In this paper, we have illuminated a critical gap between the capabilities of modern AI and the demands of real-world assistive navigation. Through the simple, yet profound, *Escalator Problem*, we demonstrated that state-of-the-art MLLMs suffer from *Implicit Motion Blindness*—a fundamental inability to perceive low-signal, continuous motion. We have argued that this is not an isolated flaw, but an inevitable consequence of the dominant frame-sampling paradigm, which sacrifices the essence of temporal dynamics for computational efficiency. This limitation poses a significant barrier to building assistive tools that BVI users can truly trust with their safety and independence.

As a position paper, our primary objective was to diagnose this foundational issue and chart a new course for the community. We have formally identified and articulated the problem, analyzed how this technical deficit metastasizes into a crisis of user trust across a spectrum of everyday scenarios, and we have issued a clear call to action. We contend that true progress in AI for accessibility cannot be measured by performance on static benchmarks alone. It requires a paradigm shift: from semantic **recognition** to holistic physical **perception**.

The path forward demands that we rethink our foundational assumptions. We must develop human-centered benchmarks co-designed with the BVI community, explore hybrid architectures that integrate classical motion-aware techniques, and investigate novel sensing modalities such as event cameras. By redirecting our focus towards the physics of motion and the unwavering standard of user trust, we can begin to build the next generation of assistive AI—systems that do not just describe the world, but empower users to navigate it safely and confidently.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 2

[2] biped.ai. NOA, 2024. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[4] Envision. Envision Glasses, 2020. 2

[5] Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding?, 2025. 2

[6] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Cuong Duc Dao. The activitynet large-scale activity recognition challenge 2018 summary, 2018. 2

[7] Google. Lookout, 2019. 2

[8] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and

Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models, 2025. 2

[9] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 2

[10] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3

[12] Microsoft. Seeing AI, 2017. 2

[13] Prudhvi Naayini, Praveen Kumar Myakala, Chiranjeevi Bura, Anil Kumar Jonnalagadda, and Srikanth Kamatala. Ai-powered assistive technologies for visual impairment, 2025. 2

[14] OpenAI. Gpt-4o system card, 2024. 1

[15] Mohammadreza Salehi, Jae Sung Park, Tanush Yadav, Aditya Kusupati, Ranjay Krishna, Yejin Choi, Hannaneh Hajishirzi, and Ali Farhadi. Actionatlas: A videoqa benchmark for domain-specialized action recognition. In *Advances in Neural Information Processing Systems*, pages 137372–137402. Curran Associates, Inc., 2024. 2

[16] Yang Shi, Jiaheng Liu, Yushuo Guan, Zhenhua Wu, Yuanxing Zhang, Zihao Wang, Weihong Lin, Jingyun Hua, Zekun Wang, Xinlong Chen, Bohan Zeng, Wentao Zhang, Fuzheng Zhang, Wenjing Yang, and Di Zhang. Mavors: Multi-granularity video representation for multimodal large language model, 2025. 2

[17] Chathurika S. Silva and Prasad Wimalaratne. Navigation framework for blind and visually impaired persons based on sensor fusion, 2025. 2

[18] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988. 5

[19] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, pages 402–419, Cham, 2020. Springer International Publishing. 7

[20] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 2

[21] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024.

[22] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025. 2

[23] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. 1

[24] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2