Are We There Yet? Exploring the Capabilities of MLLMs in Assistive AI Applications*

Shayon Dasgupta
IIT BHU
India

Avijit Dasgupta†
CVIT, IIIT Hyderabad
India

C. V. Jawahar
CVIT, IIIT Hyderabad
India

Abstract

Multimodal Large Language Models (MLLMs) have redefined visual understanding by combining vision encoders with large-scale language models. This unified architecture enables strong performance on tasks like image captioning, visual question answering, and multimodal dialogue, often in zero- and few-shot settings. Their general-purpose capabilities and flexible interfaces make MLLMs a promising foundation for real-world vision-language applications.

Assistive AI aims to help users interact with their environments through natural language. These scenarios demand robust visual recognition, contextual reasoning, and multilingual comprehension—capabilities that MLLMs are believed to offer. However, their effectiveness in assistive settings remains to be fully understood.

In this work, we explore whether MLLMs can support Assistive AI by evaluating state-of-the-art models on real-world tasks: recognizing everyday objects like currency, answering questions based on scene text, and reading visually presented content across multiple languages. To this end, we developed a system, NetraLink, using a head-mounted GoPro to capture real-world egocentric data, and collected a benchmark covering these assistive scenarios. Our findings provide a comprehensive diagnostic of current MLLMs, highlighting their strengths and limitations in enabling assistive technologies grounded in visual perception and language interaction.

CCS Concepts

• **Computing methodologies** → **Scene understanding**; *Natural language generation*; *Speech recognition.*

Keywords

Assistive Technology, MLLMs, Scene Understanding

ACM Reference Format:

Shayon Dasgupta, Avijit Dasgupta, and C. V. Jawahar. 2025. Are We There Yet? Exploring the Capabilities of MLLMs in Assistive AI Applications. In *Proceedings of 16th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'25).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3774521.3774575

*Produces the permission block, and copyright information

†Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ICVGIP'25, Mandi, India

© 2025 Copyright held by the owner/author(s).

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/3774521.3774575

1 Introduction

Assistive AI systems aim to support individuals in navigating the world more independently, particularly those with visual, cognitive, or physical impairments. These technologies seek to provide real-time understanding of the user's environment and communicate relevant information in an accessible manner. To be effective, Assistive AI must combine robust perception, high-level reasoning, and natural interaction—enabling capabilities such as object recognition, spatial understanding, and task-relevant feedback. Notably, India alone is home to over 4.95 million people with visual impairment and nearly 35 million individuals with disabilities and 0.24 million blind children overall [27], underscoring the widespread need for scalable and inclusive assistive solutions.

Human visual attention is naturally drawn to high-level semantic elements, and prior eye-tracking studies have shown that text is one of the most consistently fixated regions in natural scenes—second only to faces [7, 8, 42]. In particular, Cerf et al. [8] found that during free-viewing, people looked at text regions over 11× more often than control areas matched for size and location, underscoring the inherent salience of text in our visual environment. Importantly, participants struggled to avoid looking at text even when instructed to do so, suggesting that textual content is not only informative but also difficult to ignore. This has profound implications for Assistive AI: for individuals with visual or cognitive impairments, much of the crucial information encountered in daily life—signs, instructions, labels, or digital content—is locked behind text. Real-time text understanding can help users reconnect with their surroundings.

Motivated by the importance of visually presented text in human attention and daily functioning, we turn to recent advances in MLLMs [24, 25, 40], which have dramatically expanded the capabilities of modern vision systems by enabling them to interpret and reason over visual content through natural language. These models combine the powerful language understanding abilities of Large Language Models (LLMs) with high-capacity visual encoders, allowing them to process complex visual scenes and generate rich, contextually grounded responses. As a result, MLLMs have shown impressive performance across a wide spectrum of tasks, including image captioning [30], object detection [44], segmentation [19].

Among the most promising application areas for MLLMs is Assistive AI, which demands real-time visual understanding, language-based reasoning, and seamless interaction to augment human capabilities. These systems can be transformative for individuals with disabilities, supporting tasks such as navigating unfamiliar environments, reading text, identifying objects, and following instructions. This shift in vision-language modeling opens exciting possibilities for building more adaptive and user-aligned AI. Yet, it remains unclear whether current MLLMs are ready for real-world assistive use, where robustness, relevance, and responsiveness are essential.

To explore the real-world applicability of MLLMs in Assistive AI, we develop a prototype system, NetraLink, that simulates an egocentric assistive setup using a head-mounted GoPro[1] camera. The system is capable of capturing images on user instruction and routing them through an MLLM / LLM remotely hosted on a laptop. This flexible, offline architecture enables real-time, interactive tasks without reliance on cloud infrastructure—making it suitable for practical, privacy-aware deployments. Using this setup, we systematically study the capabilities and limitations of existing MLLMs across several assistive scenarios common to our day-to-day lives.

We categorize our evaluation into three core use cases: (a) Blind assistance – tasks like currency recognition, navigation, and scene-text question answering in complex environments; (b) Multilingual understanding – reading menu cards in Indic and English, crucial for accessibility in multilingual regions; (c) Reading comprehension – where the system observes a user reading a book over an extended period and is later queried to recall or summarize information, testing long-range grounding and memory. For each, we curate and annotate a custom dataset capturing egocentric, real-world interactions. We benchmark multiple open-source MLLMs to assess their alignment with assistive needs, analyzing strengths, failure modes, and practical readiness for everyday deployment.

In summary, this work presents the first systematic study of MLLMs for egocentric Assistive AI applications. Our key contributions are as follows:

- We develop a head-mounted assistive AI system, NetraLink, to simulate real-world data capture and interaction in egocentric assistive scenarios for evaluating the capabilities of MLLMs.
- We define five practical assistive tasks—scene-text recognition, currency-note recognition, navigation assistance, multilingual menu reading and interpretation, and long-form content reading.
- We collect and annotate a novel egocentric dataset spanning diverse real-world assistive scenarios.
- We benchmark multiple MLLMs on our dataset to evaluate their effectiveness in assistive contexts and highlight their failure modes.

2 Related Works

Egocentric Vision. Egocentric vision—where visual data is captured from a first-person viewpoint—is gaining traction in the computer vision community [10, 13, 14]. This perspective offers unique advantages for understanding user interactions, capturing contextual cues like gaze, motion, and hand-object contact. These characteristics make egocentric data particularly valuable for Assistive AI applications, as they closely mirror how users perceive and experience their surroundings. While large-scale benchmarks such as Ego4D [13] and Ego-Exo4D [14] have catalyzed progress on tasks like action recognition and object tracking, they remain limited in their coverage of practical assistive scenarios—especially those involving fine-grained visual reasoning, multilingual text understanding, and real-time human-in-the-loop interaction. To address this gap, we develop a new egocentric dataset explicitly focused on everyday assistive use cases grounded in real-world visual challenges.

Multi-modal Large Language Models. MLLMs [24, 25, 40, 45] extend LLMs with the ability to perceive, reason, and generate across multiple modalities. They typically comprise a modality encoder, a pre-trained LLM, and a projector that aligns representations across modalities. A key milestone in this direction was CLIP [32], which introduced vision–language pretraining to learn a shared representation space for images and text, laying the groundwork for unified multimodal reasoning. Building on this foundation, recent MLLMs have achieved strong results in diverse tasks such as image captioning [30], object detection [44], segmentation [19], and visual question answering [9]. By integrating perception and language understanding within a single framework, MLLMs eliminate the need for highly specialized task-specific architectures, offering a scalable, instruction-driven approach to vision problems.

Despite these advances, it remains uncertain how close we are to realizing general-purpose personal AI assistants that can operate robustly in day-to-day life. While many individual tasks have seen substantial gains, comprehensive evaluations of MLLMs in realistic assistive scenarios—particularly those involving egocentric visual data—are still lacking. In this work, we address this gap by systematically assessing state-of-the-art MLLMs on a set of everyday assistive use cases.

Text Recognition and Understanding. Text is a rich source of semantic information in our visual world, appearing on signs, screens, product labels, and documents. Accurate recognition and interpretation are crucial for tasks like navigation, reading instructions, and object identification—especially for individuals with visual impairments. Early OCR systems such as Tesseract [37] and EAST [46] have evolved into deep learning–based models like CRNN [34], ASTER [35], CRAFT [3], and DBNet [22], while document understanding approaches such as LayoutLMv3 [16], Donut [18], and DocFormer [2] enable reasoning over structured layouts. More recently, multi-modal large language models (MLLMs) like BLIP-2 [21], Kosmos-2 [31], and GPT-4V [2] have shown strong zero-shot text-grounded reasoning.

Benchmarks such as DocVQA [28], SceneText-VQA [5], COCO-Text [41], and TextVQA [36] have advanced the field, but mostly focus on static or curated images. These settings fail to capture the challenges of egocentric vision—where text may be occluded, motion-blurred, or context-dependent. Moreover, traditional OCR with rule-based NLP pipelines may yield higher accuracy and lower latency for simple text extraction, they are brittle and fail on open-ended, context-dependent queries. In contrast, MLLMs provide a unified, end-to-end approach that couples perception with language reasoning—enabling flexible, conversational, and intelligent interaction. Our work fills this gap by evaluating recognition and understanding in real-world egocentric assistive scenarios using both OCR- and MLLM-based methods. While Karamolegkou *et al.* [17] take a first step, a thorough assessment of MLLMs in multilingual, noisy, and navigation-oriented settings has been missing—precisely what we provide in this paper.

[1] https://gppro.in/product/gopro-hero10-black/

[2] https://openai.com/contributions/gpt-4v/

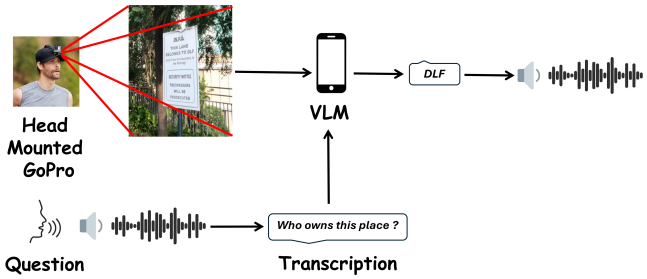

Figure 1: Overview of our Assistive AI system, NetraLink.

3 NetraLink: System Overview

Fig. 1 illustrates the end-to-end components of our egocentric assistive AI prototype. The system is organized into five tightly integrated yet modular components: (1) Audio Front-end, (2) Speech-to-Text, (3) Image Acquisition, (4) Vision–Language Inference, and (5) Response Generation and Logging. This modularity not only clarifies the dataflow but also enables independent development, testing, and future substitution of individual modules to accommodate advances in each subfield.

Audio Front-end. This module continuously monitors ambient sound and detects when the user explicitly invokes the system. We use a wake-word detector on short, overlapping audio snippets (e.g., 3 s), combining lightweight frame-based energy analysis with a voice-activity detector (VAD)[3]. This two-stage design reduces false triggers from background noise or unrelated speech, conserving computation and protecting privacy. When the wake-word is recognized, the module signals readiness and switches to active query capture, again using the same VAD in streaming mode: it buffers audio frames until it observes sustained silence beyond a configurable threshold (e.g., 1 s of non-speech). This silence-based segmentation adapts to speaking rate and ambient conditions, capturing the full utterance without manual start/stop. A small circular buffer is maintained internally to avoid clipping the start of the query, improving robustness in real-world use.

Speech-to-Text. Once the audio front-end produces a segmented utterance, it is sent to the Speech-To-Text (STT) module for transcription using a locally hosted or on-device recognizer invoked via a simple command-line interface. The STT engine applies a state-of-the-art acoustic and language model stack to perform noise reduction, phoneme decoding, and punctuation insertion, converting waveforms into text. The output is then normalized—tokenized, lower-cased, and trimmed—before being passed to the next stage. Because STT runs as a separate subprocess, we can swap or update transcription backends without affecting capture or reasoning modules. In practice, our STT module is robust, achieving an average word error rate (WER) of approximately 7% in noisy environments across 50 samples.

[3]https://github.com/wiseman/py-webrtcvad

Image Acquisition. After transcription, the system immediately captures the visual context relevant to the user's request. A head-mounted, user-facing camera provides a continuous live view but records only a single high-resolution image when triggered. We interface with it through a lightweight API that exposes a simple 'capture()' call. Each frame is time-stamped and given a unique identifier based on the capture epoch, then stored in a structured directory hierarchy organized by date and session. This keeps the camera unobtrusive—remaining in standby until explicitly activated—while conserving battery life and limiting unnecessary data capture. Storing each image alongside its corresponding transcript also provides precise cross-modal alignment for error analysis, user studies, and future module retraining.

Vision-Language Inference. At the core of the system is the vision–language inference engine. We build a single multimodal prompt by combining the user's transcribed query with a reference to the captured image, following a template such as:

`"Image Context: [image_ID] User Query: [transcribed_text]"`

This prompt conditions both the visual encoder and the language model. The engine runs as a standalone command-line process, with arguments specifying the model checkpoint, image path, prompt string, and generation parameters (e.g., maximum tokens, temperature). We measure latency from process launch to response completion for profiling and optimization. Decoupling inference into its own executable cleanly separates concerns and lets us swap architectures, upgrade to newer foundation models, or tune hyperparameters without modifying the audio front-end.

Response Generation and Logging. The raw text from the vision–language engine is first cleaned to remove system artifacts (e.g., special tokens) and then logged to persistent storage. For our prototype, we use a simple spreadsheet (Excel), appending rows with image ID, user query, model response, inference latency, and timestamp. This structured log supports both quantitative benchmarking and qualitative inspection. For real-time feedback, the cleaned response is sent to the text-to-speech (TTS) module, which synthesizes a natural-sounding waveform. The TTS engine is also invoked as an external process, with configurable voice and language settings, and the resulting audio is played through the headset speakers. We store both the intermediate text and the generated audio in a dedicated directory, enabling later user studies and ensuring full reproducibility.

Design Highlights and Extensibility. The system is built from modular components with clean interfaces, making it easy to swap wake-word detectors, STT engines, or inference backends. All processing runs locally to preserve privacy, with no cloud communication. Wake-word gating and silence-based segmentation keep the system efficient and responsive, even on limited hardware. A single configuration file controls key parameters (e.g., VAD thresholds, wake-word sensitivity, inference settings), simplifying tuning and deployment. Note that system-level optimizations such as model compression, knowledge distillation, and efficient architectures are beyond the scope of this paper, as our primary objective is to first establish and analyze the fundamental limitations of MLLMs in assistive scenarios. We consider these optimisations as natural next steps toward deployable systems.

Task	Sample Size	Source of data
1	8025 images	133 min of egocentric video
2	150 images	Banknotes from 5 countries
3	3 environments	133 min of egocentric videos
4	72 menu images	In-the-wild English, Bengali, Hindi menu photos
5	37 page-level images	62 min of egocentric story-book reading

Table 1: Summary of the five tasks in our dataset. Extending the dataset with additional tasks is part of future work.

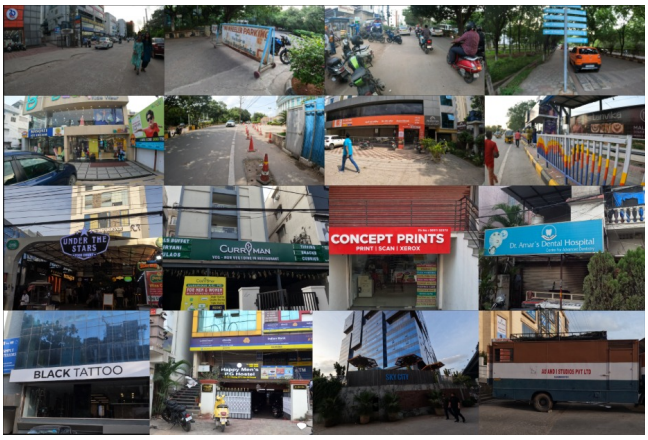

Figure 2: Examples of egocentric images used to evaluate MLLMs for scene-text recognition, captured using our head-mounted assistive AI system across diverse urban environments—including sidewalks, pedestrian crossings, market areas, and residential streets. These images feature varied instances of scene text such as signs, storefronts, advertisements, and labels, which are vital for real-world understanding and navigation.

4 Benchmark Setup

In this section, we outline *five* practical tasks of the developed head-mounted assistive AI system—each centered around real-world text understanding tasks as motivated in Section 1—and describe the associated datasets used to evaluate these capabilities. We provide an overview of all tasks and dataset statistics in Table 1.

Globally, over 2.2 billion people experience near or distance vision impairment [6, 12], including an estimated 43.3 million who are blind [29]. This widespread prevalence highlights the critical need for effective assistive technologies that can support blind and visually impaired individuals in navigating daily challenges. To help address these needs, we focus on three key tasks as follows -

Task 1: Scene-Text Recognition. Text is one of the most important visual elements in our daily lives [7, 8, 42]. Studies show that people consistently focus on text in natural scenes—often even more than on other visual cues—due to its high-level semantic value. Whether it's a street sign, product label, or digital interface, written text helps individuals interpret and interact with their environment.

Figure 3: Examples of currency-note images used to evaluate MLLMs across five countries: India, the USA, the UK, Europe, and Hungary. Each note is paired with multiple question types—such as "What is the value of this note?"—to assess the model's ability to recognize denominations and reason about visual financial cues under diverse layouts, languages, and designs.

For blind individuals, this information remains largely inaccessible, making navigation, object recognition, and understanding instructions challenging. Real-time scene-text recognition can bridge this gap, enhancing independence and situational awareness.

We frame this as a visual question answering (VQA) task, where both questions and answers are in natural language. Using our NetraLink, we captured 133.31 minutes of egocentric video across varied urban settings with a head-mounted GoPro, sampling frames at 1 FPS to obtain 8,025 images. Example questions include "What does the text on the sign say?" or "Which shop is this?"—assessing a model's ability to detect, read, and reason about text in context (see Fig. 2).

Task 2: Currency-Note Recognition. The goal of this task is to assess a model's ability to answer questions about paper currency across diverse designs. For blind individuals, banknote identification is challenging due to the lack of consistent tactile markers, making transactions—such as paying, receiving change, or verifying denominations—difficult and error-prone. Real-time currency recognition can enhance financial independence and confidence in economic interactions.

We frame this as a VQA task, with both questions and answers in natural language. Our dataset contains 150 images of banknotes from five countries—India, the USA, the UK, Europe, and Hungary—captured to represent varied designs and denominations. For each of the 30 images per country, we created four question types (e.g., "What is the origin of this note?") to evaluate recognition accuracy and reasoning over currency-specific visual features (see Fig. 3). This design enables cross-cultural, denomination-aware evaluation and encourages robustness in real-world financial scenarios.

Task 3: Navigation Assistance. This task evaluates a model's ability to interpret egocentric scenes for real-time, language-guided navigation. Moving through unfamiliar or dynamic environments is particularly challenging for visually impaired individuals, as critical cues—such as signs, crossings, and pathways—are primarily visual. While canes and GPS-based audio tools aid mobility, they provide limited spatial and semantic context. Vision-language systems can bridge this gap by offering context-aware guidance, improving both safety and independence.

Figure 4: Sample egocentric images captured in varied outdoor environments—including a university campus, residential neighborhood, and commercial market—used to evaluate the navigation capabilities of MLLMs in dynamic, real-world settings.

The evaluation is posed as a natural language question–answering problem grounded in visual input. Using our NetraLink, we recorded egocentric video in three outdoor settings: a university campus, a residential neighborhood, and a commercial market. From this footage, we extracted high-resolution frames covering diverse mobility scenarios. A subset was annotated with navigation-focused questions such as "What is in front of me?", "Is there a crossing ahead?", and "Can I turn right here?" to assess spatial reasoning and scene interpretation (see Fig. 4).

Understanding visual information across multiple languages is difficult for both people with visual impairments and the general population. Public materials like menus, signboards, and official documents often mix scripts and languages, causing confusion even for locals. This makes robust multilingual visual understanding a key requirement for assistive and general-purpose AI, motivating the following task.

Task 4: Multilingual Menu Card Reading and Interpretation. Restaurant menus are a prototypical example of complex multilingual content, with diverse layouts, decorative fonts, and mixed-language text (e.g., dish names in Hindi or Bengali with English descriptions). For someone who cannot clearly see the menu or does not know one of the scripts, this becomes a major barrier. An assistive AI system should visually parse the menu, understand its content across languages, and answer natural-language questions.

Using our NetraLink platform, we build a head-mounted multimodal menu-card VQA setup. The task is to answer natural-language questions about menu content by jointly leveraging visual and textual cues across languages. We curate 72 real-world menu images (Fig. 5): 18 in English, 28 in Bengali, and 26 in Hindi, covering single-language, bilingual, and mixed-script layouts. Captured under varied conditions, they vary in fonts, image quality, and background clutter. Each image is annotated with aligned question–answer pairs, enabling systematic evaluation of a model's ability to extract, interpret, and reason over multilingual text in realistic scenarios.

Understanding long-form content is a challenge that extends beyond assistive contexts, affecting anyone who needs to follow a narrative, track entities, or integrate information over time. Books, reports, and other extended texts require readers to maintain continuity across multiple observations rather than extracting meaning

Figure 5: Sample images of restaurant menu cards in Hindi, Bengali, and English used to evaluate multilingual text understanding. The dataset includes printed menus in varied layouts and scripts, reflecting real-world diversity in visual and linguistic presentation.

Figure 6: Sample page-level images from two printed storybooks captured using our head-mounted assistive AI system. These images are used to evaluate long-form narrative understanding through contextual question answering on characters, relationships, and plot elements.

from isolated snippets. For AI systems, mastering this skill is critical for applications in education, storytelling, summarization, and real-world reading assistance—enabling richer, context-aware interaction with extended materials.

Task 5: Reading Long-form Content. Storybooks are a natural example of long-form visual text that demands continuity tracking and deeper reasoning. Unlike signage or short passages, where meaning can be grasped from a single frame, narratives in printed books require models to recognize and integrate information across multiple pages. This involves accurate OCR, entity and relationship tracking, and maintaining plot context—skills essential for delivering meaningful, context-aware answers to reader queries.

To evaluate this capability, we curated a dataset by recording 62 minutes of egocentric video while reading two printed storybooks using our head-mounted assistive AI system, NetraLink. From these recordings, we extracted 37 page-level images under realistic reading conditions, which include challenges such as varied lighting, page curvature, partial occlusions from hands, and non-uniform

alignment. Each page is annotated with narrative-focused question–answer pairs in natural language, requiring reasoning over one or more pages. The questions span character identification, relationship mapping, and plot development (e.g., "Who is the miserly old man in the first story?", "Why does Vrindavan Kundu quarrel with his father?", "What medical recommendation caused the dispute?"). This benchmark measures a model's ability to couple OCR with long-range, context-aware reasoning, advancing both assistive story-based QA and general-purpose narrative comprehension. Some samples are shown in Fig. 6.

5 Evaluation and Analysis

We evaluate a set of state-of-the-art MLLMs—including Gemma-3-4B [38], Gemma-3-12B [38], Gemma-3-27B [38], LLaVA [26], LLaVA-LLaMA3 [33], LLaVA-Phi3-3.8B [33], Qwen2.5-VL-7B [4], and BakLLaVA [23]—to analyze their effectiveness across five assistive AI applications (**Task 1-5**) proposed in this work. Each task is cast as a question answering (QA) problem to ensure a consistent and unified evaluation framework across heterogeneous input types. This design enables a joint assessment of both visual recognition and language reasoning abilities. In line with previous works [5, 28], we employ the Average Normalized Levenshtein Similarity (ANLS) score as our primary evaluation metric to quantify response accuracy. We also evaluate the LLM-as-a-judge [20] to evaluate the correctness of the answers. In addition to the ANLS score, we also recorded the average inference latency for each model to assess their practicality in real-time assistive applications, highlighting trade-offs between accuracy and responsiveness.

Table 2: Quantitative results for Task 1: Average latency and ANLS for each model across all scene-text queries, with the best and second-best performers highlighted in bold and underline, respectively.

Model	Latency (s)	ANLS	LLM
LLaVA [26]	3.01	0.32	0.51
Gemma-3-27B [38]	19.32	0.35	0.78
LLaVA–Phi3-3.8B [33]	**1.95**	0.36	0.45
Gemma-3-12B [38]	15.73	0.42	0.77
Gemma-3-4B [38]	13.37	0.44	0.71
BakLLaVA [23]	<u>2.17</u>	0.47	0.43
LLaVA–LLaMA3 [33]	2.65	0.50	0.50
Qwen2.5-VL-7B [4]	26.13	<u>0.57</u>	0.76
FastVLM-7B [39]	14.88	**0.66**	0.85

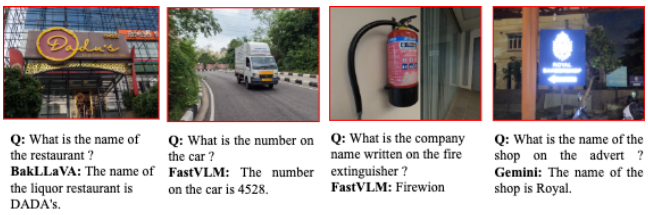

Figure 7: Qualitative examples of failure cases for scene-text recognition.

Task 1: Scene-Text Recognition. Text is integral to daily life — appearing on signs, storefronts, menus, and instructions—making its recognition vital for assistive technologies. To show this, we processed 8,025 video frames (113 minutes of footage) using the `IndicPhotoOCR`[4] engine on a mid-range ARM CPU, recording word counts and variability under cluttered and sparse-text scenes. The system recognized 119,791 words (14.93/frame), demonstrating the significant volume of textual information available in everyday environments and the importance of accurate extraction for real-world assistive applications.

To assess the suitability of modern MLLMs for assistive reading in real-world settings, we evaluated nine models on a benchmark of natural images with printed and handwritten text under cluttered, low-contrast, and occluded conditions. FastVLM-7B [39] achieved the highest ANLS (0.66), followed by Qwen2.5-VL-7B [4] (0.57), while Gemma-3-4B [38] (0.44) outperformed larger variants, showing that parameter scale alone does not ensure fine-grained text understanding. For latency-sensitive scenarios, LLaVA–Phi3-3.8B [33] (0.36 ANLS, 1.95s) and BakLLaVA [23] (0.47 ANLS, 2.17s) provided the best accuracy–speed balance, underscoring the importance of vision–language alignment and task-specific grounding over model size. This is consistent with typical scaling trends, inference latency increases with LLM parameter count. FastVLM-7B [39] likewise attains the top LLM-as-a-judge [20] score, underscoring its strong text understanding and reasoning capabilities.

Our analysis of failure cases (see Fig. 7) reveals four recurring challenges in scene-text recognition with MLLMs: (a) non-standard fonts, (b) small text size, (c) cluttered text, and (d) challenging lighting. Non-standard or stylized fonts, such as the cursive script on a restaurant sign, caused the model to misread "Dadu's" as "DADA's," highlighting a lack of robustness to typography underrepresented in training data. Small text relative to image resolution often went undetected, as in the case of a car license plate "TS 08 T 4528" being read only as "4528." In cluttered settings, dense arrangements of instructional or decorative text with varying fonts, colors, and sizes obscured the target words—for example, the fire extinguisher brand "Firewin" was recognized as "Firewion" due to surrounding label content. Lighting conditions further impacted performance: in a night-time scene, high exposure and blooming from a glowing LED sign rendered "Barbershop" illegible, with only the more prominent "Royal" partially recognized. These examples underscore the difficulty of achieving robust performance under the diverse and imperfect visual conditions.

Task 2: Currency Recognition. Each currency image was evaluated using four standardized prompts targeting distinct aspects of currency understanding: identifying the note's value, determining its denomination and origin, describing its visual elements, and extracting any visible serial numbers or security features. As shown in Tab. 3, LLaVA-LLaMA3 [33] achieved the highest ANLS (0.2787), benefiting from strong multimodal alignment. Gemma-3-4B [38] and Qwen2.5-VL-7B [4] ranked next, with Gemma-3-4B [38] performing well despite its smaller size and Qwen's robustness aided by fine-grained vision encoding and multilingual training. Increasing model size within the same family did not consistently improve results, underscoring that architectural alignment and task-specific

[4]https://github.com/Bhashini-IITJ/IndicPhotoOCR

Table 3: Quantitative results for Task 2: Average latency, and ANLS for each model over all currency-note queries. The best and second-best performing models are highlighted using bold and underline, respectively.

Model	Latency (s)	ANLS	LLM
BakLLaVA [23]	**0.92**	0.16	0.28
FastVLM-7B [39]	15.02	0.17	0.37
LLaVA [26]	3.00	0.22	0.32
Gemma-3-27B [38]	16.03	0.24	0.98
Gemma-3-12B [38]	8.48	0.24	0.58
LLaVA–Phi3-3.8B [33]	<u>1.00</u>	0.24	0.28
Qwen2.5-VL-7B [4]	9.52	0.27	0.44
Gemma-3-4B [38]	7.46	<u>0.27</u>	0.40
LLaVA–LLaMA3 [33]	2.93	**0.28**	0.30

visual grounding matter more than parameter count for fine-grained currency recognition.

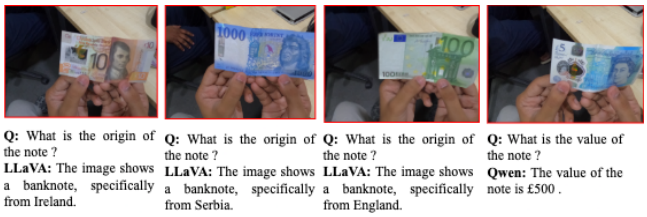

Figure 8: Qualitative examples of failure cases for currency recognition.

In currency recognition (see Fig. 8), MLLMs often falter at fine-grained classification—misidentifying denomination or origin even when correctly detecting a banknote. For instance, in Fig. 8, a Hungarian note was attributed to Serbia after the model failed to read "MAGYAR NEMZETI BANK" and instead may have guessed from visually similar notes in its training data. Folded or partially obscured notes can hide denomination markers, while blur from motion or poor focus reduces legibility of security details and serial numbers. These cases show the models' dependence on clear, unobstructed text and their limited robustness to worn, folded, or poorly captured currency.

In terms of latency, BakLLaVA [23] offers the fastest average inference at 0.92s, followed by LLaVA-Phi3-3.8B [33] (1.83s) and LLaVA-LLaMA3 [33] (2.93s), making them appealing for real-time applications. In contrast, larger models like Gemma-3-27B [38] (16.03s) and Qwen2.5-VL-7B [4] (9.52s) exhibit significantly higher inference times, highlighting a trade-off between accuracy and responsiveness that must be considered in deployable assistive systems.

Task 3: Navigation Assistance. We also assess the directional understanding of MLLMs. To do this, we extracted a set of real-world navigation routes, as detailed in Section 4. Fig. 9 presents one such qualitative example, illustrating how MLLMs interpret spatial cues and directional references within egocentric scenes. Our observations show that MLLMs can effectively guide users using navigation signs, provided the text is reliably detected and read—highlighting

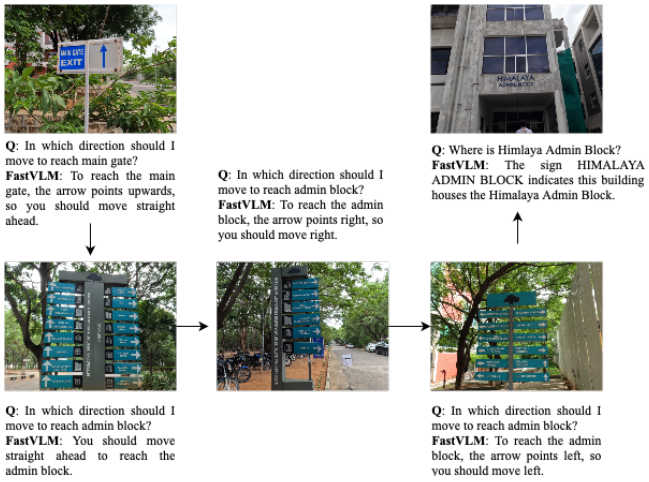

Figure 9: Example output from FastVLM [39] for a navigation query involving the destination "Himalaya Admin Block." The model correctly interprets scene elements like buildings and signage.

that overcoming the scene-text recognition challenges in Task 1 is key to unlocking their full potential in navigation assistance.

Table 4: Quantitative results for Task 4: Average latency, and ANLS for each model for Understanding Indic Languages. The best and second-best performing models are highlighted using bold and underline, respectively.

Model	English		Hindi		Bengali		Avg Latency (s)
	LLM	ANLS	LLM	ANLS	LLM	ANLS	
Gemma-3-4B [38]	0.69	0.47	0.42	0.36	0.51	**0.43**	4.86
Gemma-3-12B [38]	0.70	0.45	0.57	<u>0.37</u>	0.67	0.37	6.11
Gemma-3-27B [38]	0.96	0.48	0.83	**0.40**	0.93	<u>0.41</u>	7.82
LLaVA [26]	0.04	0.17	0.00	0.13	0.06	0.12	2.68
LLaVA–LLaMA3 [33]	0.22	<u>0.55</u>	0.05	0.28	0.03	0.27	1.54
LLaVA–Phi3–3.8B [33]	0.11	0.25	0.06	0.19	0.03	0.16	1.43
Qwen2.5–VL–7B [4]	0.19	**0.59**	-	-	-	-	-
BakLLaVA [23]	0.11	0.11	0.03	0.12	0.01	0.06	0.93
FastVLM-7B [39]	0.12	0.18	0.12	0.16	0.04	0.09	13.42

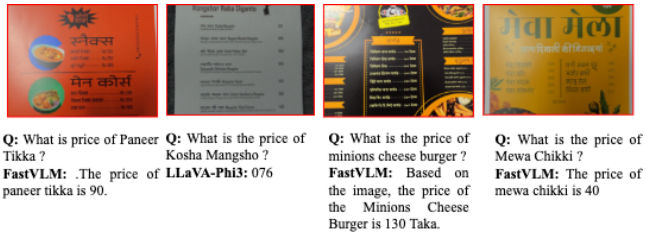

Figure 10: Qualitative examples of failure cases for multilingual menu card reading and interpretation.

Task 4: Multilingual Menu Card Reading and Interpretation. Table 4 summarizes the performance of various MLLMs on understanding restaurant menu cards across English, Hindi, and Bengali. Notably, among all evaluated models, only the Gemma family was explicitly trained with support for Indic languages; the others

were primarily trained on English and high-resource languages. Despite this, Qwen2.5-VL-7B [4] achieves the highest ANLS score on English menus (0.5588), followed closely by LLaVA-LLaMA3 (0.5460) and Gemma-3-27B (0.4807). For Hindi, Gemma-3-27B [38] outperforms all other models with an ANLS of 0.4035, with Gemma-3-12B [38] (0.3705) as the second-best. Surprisingly, for Bengali, the smaller Gemma-3-4B [38] performs best (0.4249), outperforming its larger siblings. This result suggests that multilingual pretraining in Gemma [38] helps significantly, especially for scripts and content not represented in the training distributions of other models. Due to the generation of non-decodable or illegal Unicode characters by Qwen2.5-VL-7B [4] in Indic language outputs, we were unable to include its results in the Hindi and Bengali evaluations.

In the menu card example (refer Fig. 10), the model struggles to accurately identify and associate dish names with their corresponding prices due to a combination of challenges: visually complex layouts with multiple columns, varying background colors, suboptimal lighting or exposure, and the simultaneous presence of multiple languages. Such conditions can confuse text localization, hinder script recognition, and break the logical association between related text elements. These limitations significantly reduce the robustness of MLLMs for real-world menu card understanding, especially in visually cluttered and multilingual environments. Gemma-3-27B [38] attains the highest LLM-as-a-judge performance [20], with accuracies of 96.29%, 83.11%, and 93.13% for English, Hindi, and Bengali, respectively.

In terms of latency, lightweight models such as BakLLaVA [23] and LLaVA-Phi3-3.8B [33] demonstrate the fastest inference across all languages, with BakLLaVA [23] consistently achieving sub-1s latency, making it ideal for low-latency applications. However, this speed comes with a trade-off BakLLaVA [23] shows very low ANLS scores, reflecting weak text recognition capabilities. While Qwen2.5-VL-7B [4] achieves high accuracy on English, it lacks multilingual capacity and was not evaluated on Hindi or Bengali. Interestingly, FastVLM-7B [39], despite its branding, exhibits significantly higher latency (12–14s) while delivering only moderate accuracy, suggesting that its architecture is optimized for different tasks. Overall, the results highlight that Indic-language performance is highly dependent on pretraining exposure, and that models not trained on regional scripts suffer both in accuracy and consistency—underscoring the value of multilingual pretraining for real-world assistive AI deployments in linguistically diverse regions.

Task 5: Reading Long-form Content We evaluated four MLLMs — DeepSeek-R1-14B [15], LLaMA3.1-8B [11], Phi-4-14B [1], and Qwen3-14B [43] — measuring inference latency and ANLS (Table 5). LLaMA3.1-8B [11] [15] achieved the best balance with the highest ANLS (0.1278) and low latency (13.34s). DeepSeek-R1-14B [15] followed in accuracy but was slower (56.64s). Qwen3-14B [43] and Phi-4-14B [1] underperformed in accuracy and speed, likely due to limitations in long-context reasoning. These results highlight the need for models that are both context-aware and latency-efficient for assistive reading tasks: average latency and ANLS for each model on the Book Reading QA task. Latency reflects the time to process a single query, showing large differences across models—from 13.34s for LLaMA3.1-8B [11] to over 130s for Qwen3-14B [43]—highlighting the trade-off between speed and accuracy in real-time reading assistance. The LLM scores are relatively low overall (0.22–0.29), underscoring the difficulty of the task. This suggests that, despite similar ANLS values, models differ in how often their responses are judged semantically adequate by another LLM.

Table 5: Quantitative results for Task 5: Average latency, exact-match accuracy, and ANLS for each model on the Book Reading QA task.

Model	Latency (s)	ANLS	LLM
DeepSeek-R1-14B [15]	56.64	0.12	0.28
LLaMA3.1-8B [11]	13.34	0.13	0.24
Phi-4-14B [1]	30.57	0.06	0.22
Qwen3-14B [43]	130.07	0.07	0.29

Our findings reveal two key challenges: in multilingual scenarios (Task 4), limited script exposure leads to basic recognition failures (e.g., LLaVA, Qwen) or misalignment in complex layouts (e.g., Gemma). In long-context reading (Task 5), accumulated OCR errors and weak reasoning across sequential inputs hinder models that handle single-image VQA well but fail to maintain narrative consistency across pages. We present further qualitative examples for all tasks in the supplementary.

6 Conclusion

In this work, we explored the potential of MLLMs for a spectrum of visual assistive tasks, including scene-text recognition, currency note identification, navigation assistance, multilingual menu reading, and long-form content access. Our results show that, when coupled with speech and question-answering interfaces, MLLMs offer a promising route toward AI assistants that substantially improve accessibility for people with visual impairments. Such systems can deliver real-time, context-aware information, helping bridge the gap between the visual world and non-visual perception. Our findings generalize broadly, as the tasks test core abilities—text recognition, navigation, and object understanding—while exposing real-world challenges like poor lighting, small fonts, complex layouts, multilingual text, and contextual reasoning.

At the same time, our experiments reveal clear limitations: performance degrades in complex, real-world scenarios with multi-column layouts, low light, cluttered backgrounds, and multilingual content, underscoring gaps in robustness. Future work should therefore focus on domain- and task-specific fine-tuning, as well as on designing more natural, accessible interfaces tailored to visually challenged users. Reducing latency will also be crucial, since slow responses disrupt interaction and undermine usability in time-sensitive situations. With these targeted improvements, MLLMs can progress from promising prototypes to dependable assistive technologies suitable for everyday use.

7 Acknowledgments

Shayon Dasgupta is supported through the SRFP by the Indian Academy of Sciences. Avijit Dasgupta is supported by a Google Ph.D. India Fellowship. The work is supported by MeitY, Govt. of India, through the NLTM-Bhashini project.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905* (2024).

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *ICCV*.

[3] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character Region Awareness for Text Detection. In *CVPR*.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[5] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *ICCV*.

[6] Matthew J Burton, Jacqueline Ramke, Ana Patricia Marques, Rupert RA Bourne, Nathan Congdon, Iain Jones, Brandon AM Ah Tong, Simon Arunga, Damodar Bachani, Covadonga Bascaran, et al. 2021. The lancet global health commission on global eye health: vision beyond 2020. *The Lancet Global Health* (2021).

[7] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. 2016. Where should saliency models look next?. In *ECCV*.

[8] Moran Cerf, E Paxon Frady, and Christof Koch. 2009. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision* (2009).

[9] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *CVPR*.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*.

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[12] Timothy R Fricke, Nina Tahhan, Serge Resnikoff, Eric Papas, Anthea Burnett, Suit May Ho, Thomas Naduvilath, and Kovin S Naidoo. 2018. Global prevalence of presbyopia and vision impairment from uncorrected presbyopia: systematic review, meta-analysis, and modelling. *Ophthalmology* (2018).

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*.

[14] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[16] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *ACM-MM*.

[17] Antonia Karamolegkou, Malvina Nikandrou, Georgios Pantazopoulos, Danae Sanchez Villegas, Phillip Rust, Ruchira Dhar, Daniel Hershcovich, and Anders Søgaard. 2025. Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users. *arXiv preprint arXiv:2503.22610* (2025).

[18] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-Free Document Understanding Transformer. In *ECCV*.

[19] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *CVPR*.

[20] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579* (2024).

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

[22] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time Scene Text Detection with Differentiable Binarization. In *AAAI*.

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*.

[24] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *NeurIPS* (2023).

[27] Sunny Mannava, Rishi Raj Borah, and BR Shamanna. 2022. Current estimates of the economic burden of blindness and visual impairment in India: A cost of illness study. *Indian Journal of Ophthalmology* (2022).

[28] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

[29] Vision Loss Expert Group of the Global Burden of Disease Study et al. 2024. Global estimates on the number of people blind or visually impaired by cataract: a meta-analysis from 2000 to 2020. *Eye* (2024).

[30] Ruotian Peng, Haiying He, Yake Wei, Yandong Wen, and Di Hu. 2025. Patch Matters: Training-free Fine-grained Image Caption Enhancement via Local Perception. In *CVPR*.

[31] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[33] Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad S. Khan. 2024. LLaVA++: Extending Visual Capabilities with LLaMA-3 and Phi-3. https://github.com/mbzuai-oryx/LLaVA-pp

[34] Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *PAMI* (2016).

[35] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2016. Robust Scene Text Recognition with Automatic Rectification. In *CVPR*.

[36] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. In *CVPR*.

[37] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *ICDAR*.

[38] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).

[39] Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, et al. 2025. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19769–19780.

[40] Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. 2025. FastVLM: Efficient Vision Encoding for Vision Language Models. In *CVPR*.

[41] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140* (2016).

[42] Hsueh-Cheng Wang and Marc Pomplun. 2012. The attraction of visual attention to texts in real-world scenes. *Journal of vision* (2012).

[43] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[44] Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. 2025. ROD-MLLM: Towards More Reliable Object Detection in Multimodal Large Language Models. In *CVPR*.

[45] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361* (2023).

[46] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *CVPR*.