Describe Now: An AI-Powered Video Understanding System for On-Demand Audio Descriptions

Aarya D Roy[1], L. Sumathi[2]
[1]PG scholar, Department of CSE, Government College of Technology, Coimbatore, India
[2]Assistant Professor, Department of CSE, Government College of Technology, Coimbatore, India

Abstract— Blind and low vision (BLV) people face challenges accessing visual information in video content. Existing audio description systems often involve high production costs and dependence on paid cloud APIs. This paper presents Describe Now, a user-driven real-time audio description system that enables BLV users to request AI-generated descriptions during video playback. The system integrates the open-source LLaVA vision-language model with a Python Flask backend and OpenCV-based frame extraction pipeline. Users can request concise (≤25 words) or detailed (≤100 words) descriptions through voice commands or keyboard shortcuts.The system also provides several accessibility features, including hover-based text-to-speech support, continuous voice commands, structured audio cues, keyboard-only navigation and ARIA live-region screen reader support. Evaluation was conducted using 6 test videos and 144 generated descriptions. The system achieved a combined description quality score of 97.54% using BLEU, ROUGE and METEOR metrics. All generated descriptions satisfied their word-count limits, achieving 100% compliance. In addition, the system passed all WCAG 2.1 Level AA accessibility criteria. The proposed system operates fully offline after model installation, eliminating API costs while improving user privacy.

Index Terms—Audio Description, Blind and Low Vision Accessibility, Flask, LLaVA, Vision-Language Model, Web Accessibility.

I. INTRODUCTION

Nowadays, video plays a major role in sharing information across virtually every domain — education, news, entertainment, healthcare — but still this medium is unavailable for a large population globally. According to the World Health Organization, at least 1 billion people have different forms of vision problems [WHO:2023], In a figure around 36 million people have been classified as blind and 217 million people have been classified as having moderate to severe visual problems recorded by the Vision Loss Expert Group [Bourne et al., 2017]. While closed captions have made substantial progress in addressing the needs of deaf and hard-of-hearing users, no comparable mechanism exists to describe the visual dimension of video to BLV audiences. Audio description — a supplementary narration track that conveys what is happening on screen — is the accepted solution, but the economics of professional AD production have kept it confined largely to major broadcast properties

Recent advances in vision-language models (VLMs) offer a promising path forward. Models capable of interpreting images and generating natural language descriptions can potentially automate the audio description process in real time. However, existing research either relies on expensive proprietary cloud APIs, requires pre-generation of all descriptions before viewing, or lacks the accessibility infrastructure needed for BLV users to actually interact with the system. This project addresses all three limitations.

Describe Now is a web-based video accessibility platform that integrates a locally hosted vision-language model (LLaVA) with a Flask backend and a fully accessible frontend. Users can trigger descriptions at any moment using keyboard shortcuts, voice commands or any assistive technology compatible with web standards. It is free to use, secure, working offline after completing the installation process in real time and all the AI processing happens locally.

Current approaches to video accessibility for BLV users suffer from several fundamental limitations. Traditional human-authored audio descriptions require expensive professional narrators and post-production

studios, making them economically viable only for major broadcast productions. AI-based systems that rely on cloud APIs such as GPT-4V impose per-request costs that make large-scale deployment infeasible for educational institutions and non-profit organisations. Blind and low vision users cannot control when and how to receive the descriptions frequently that's why pre-generated description fails for users agency. Most of the existing web-based video players are not designed with blind and low vision user as a primary listener, fails to support proper ARIA attributes, keyboard navigation and screen reader support.

1.1 Objectives:

1. To design and implement an audio description system that will help BLV users to get the AI-generated video description when they need.
2. Using LLaVA as the AI engine to eliminate dependency on cloud APIs.
3. The description will be provided at two levels: Concise(<=25 words) and detailed (<=100 words).
4. Implementing five novel features to access the system.
5. Building a video management system that will help to support to upload, switch and delete operation on user-interface.

II. RELATED WORK

For BLV users there are lots of research in different parameters to access the video. Cheema, Seifi and Fazli [1] introduced a user-driven audio description system using GPT-4V, presented at DIS 2025. It offers two types of description modes – concise (<=25 words) and detailed (<=100 words) and it is activated using keyboard shortcut 'C' and 'D'. A formal user study used 20 BLV participants around seven types of video to understand and improve the accuracy as compare to traditional methods. This system is completely depends on paid cloud API, pre-generated description instead of producing in real time and there are no other keyboard shortcuts besides the two keys for accessibility features.

Li et al. [2] They created a system called VideoAlly. It uses multimodal LLMs (to understand the text+image) with 42 professional audio description guidelines that will guide to GPT-4 to generate the high quality of description. They have created a dataset with 40000 annotated videos to validate the structured-prompt engineering approach. Ning et al. [3] Developed a system called SPICA, it enables interactive scene exploration with the help of spatial positioning and it inspiring to the audio cue system in this work. Aafaq et al. [4] studied evaluation metrics for video description, noting that BLEU tends to favour shorter descriptions.

Yuksel et al. [5] proposed a human-in-the-loop machine learning framework that integrating automatic caption generator with checked by human and highlighting the critical need for a fully automated pipeline. Rescribe [6] It is addressed for post-production audio description placement but its not supporting in real time. Bodi et al. [7] demonstrate that combining baseline descriptions with interactive follow-up produces higher user comprehension, which support two mode concise and detailed and designed adopted in this work. Chuang and Fazli [8] They had introduced a system called CLearViD, it is a curriculum learning-based video description model that is to training model progressively. In-short, this model adopt zero-shot prompting using a pretrained vision language model to eliminate the need for additional training.

Chang et al. [9] developed a system called WorldScribe, a system closely related to this work that provides real-time descriptions for BLV users navigating physical environments. According to this system, its relies on GPT-4V, a paid cloud model and need transmitting user environment data to external server, which is concern about cost and privacy. Quero et al. [10] showed that combining non-verbal audio signals with verbal descriptions improves the experience of visually impaired users. This design influenced the multi-tone audio cue system in the present work.

2.1 Research Gap

Based on the literature review, it is observed that no existing system combines all these features in a single application. (a) real-time on-demand Artificial Intelligence description generation, (b) zero cost with no cloud dependency, (c) full compliance with WCAG 2.1 Level AA and (d) keyboard shortcuts to interact with the system. In particular, there are no features like hover reading for user, voice command, and

programmatic audio cues in any of the ten system reviewed. This proposed Describe Now system addresses these gaps through its overall architecture, achieving a description quality score of 97.54% along with 100% word count. Description quality is assessed using standard evaluation metrics, including BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), and METEOR (Metric for Evaluation of Translation with Explicit Ordering), which measure textual similarity and semantic accuracy.

Table I Overview of Surveyed Papers

Ref	Author/Year	System	Real-Time	Free/ Local	BLV	Limitations	Relevance to This Work
[1]	Cheema et al. 2025	Describe Now: User-Driven AD	No	No	Yes	Cloud API; pre-generated; no a11y features	Primary baseline — C/D key design adopted
[2]	Li et al. 2025	VideoA11y Dataset & Framework	No	No	Yes	Cloud API; fine-tuned models lose generalizability	Validates compliant-prompt approach
[3]	Ning et al. 2024	SPICA: Spatial Audio AD	Partial	No	Yes	Manual exploration disrupts narrative flow	Interactive description paradigm; audio cues
[4]	Aafaq et al. 2019	Video Description Survey	N/A	N/A	Limited	Survey only; metrics misaligned with human judgment	Metric selection rationale; benchmark landscape
[5]	Yuksel et al. 2020	HILML for Accessibility	No	Partial	Yes	Requires human intervention; moderate accuracy	Highlights need for fully automated pipeline
[6]	Rescribe 2020	AD Post-Production Tool	No	Partial	Yes	Post-production only; depends on audio gaps	Timing and text compression strategies
[7]	Bodi et al. 2021	NarrationBot + InfoBot	Partial	Partial	Yes	Small sample; detection accuracy varies	Direct lineage to [1]; C/D mode motivation
[8]	Chuang & Fazli 2023	CLearViD Curriculum Learning	No	Yes	No	Instructional videos only; no BLV focus	Rationale for zero-shot over trained model
[9]	Chang et al. 2024	WorldScribe Real-Time AD	Yes	No	Yes	Cloud API; hallucinations; privacy concerns	Real-time AI — motivates local deployment
[10]	Quero et al. 2020	Multimodal Artwork Guide	Yes	Partial	Yes	Hardware-dependent; not scalable to video	Audio cue design inspiration

III. METHODOLOGY

This section explains the system architecture, data flow and key design of Describe Now. The proposed

system provides a fully accessible offline environment for users and it does not depend on paid API with three layer of architecture.

3.1 System Architecture

This system is organized in three tiers. The Presentation Tier runs in the browser and houses five accessibility modules (keyboard navigation, voice commands, hover reading, audio cues, and ARIA/screen reader support) alongside the HTML5 video player, description display, TTS output, and video management interface. The Application Tier is a Python Flask server (Python 3.10+, port 5000) exposing 14 REST endpoints across five Python modules. The Data Tier is a SQLite database with descriptions and sessions tables. All AI inference is handled by a local engine: Ollama running LLaVA 7B at localhost:11434. Figure 1 shows the complete architecture.

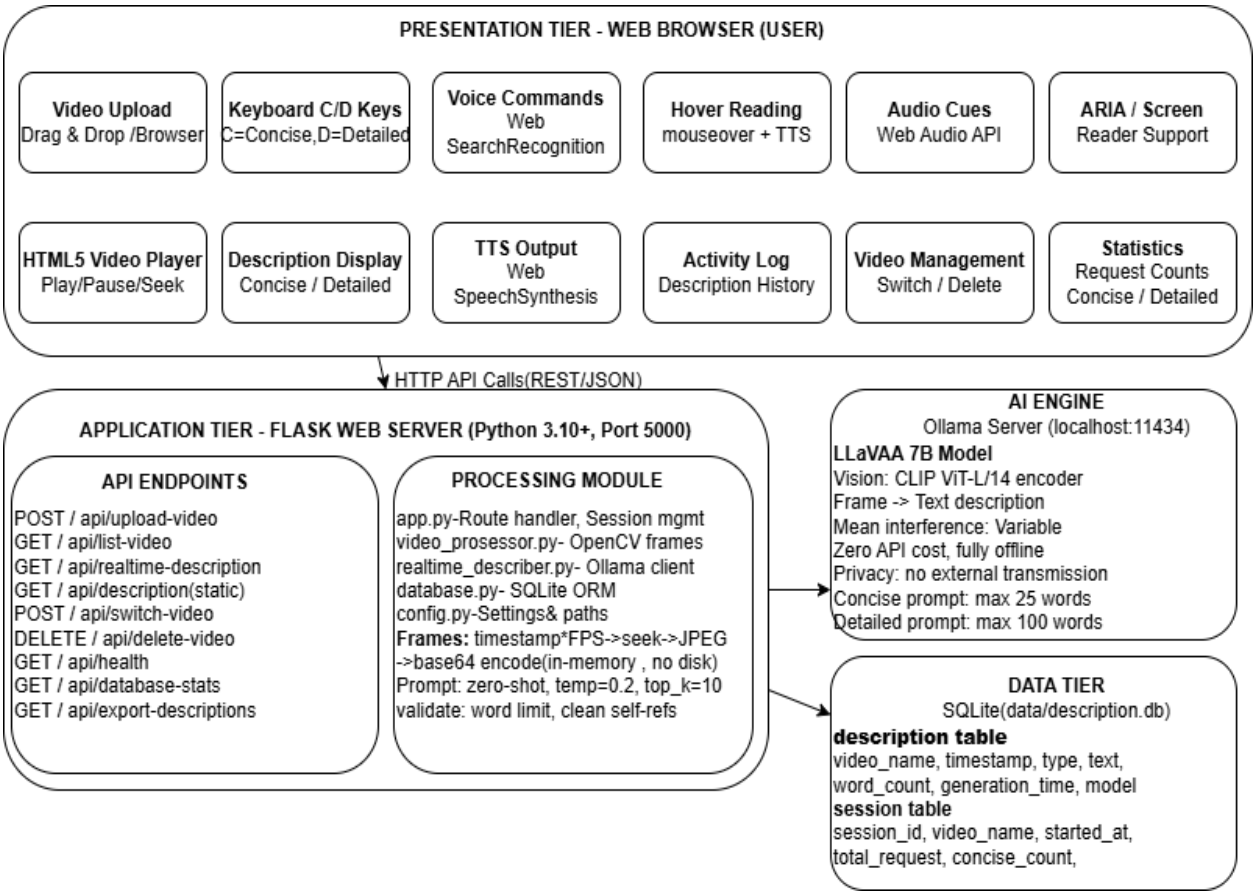

Fig. 1. System Architecture

3.2 Data Flow Pipeline

1. Every user can trigger via voice command or keyboard C/D and it passes a through 10-step pipeline.
2. JS reads currentTime, pause the video and send to GET request.
3. Flask validates and routes
4. OpenCV seeks frame at timestamp × FPS.
5. JPEG encode in-memory → base64.
6. The encoded frame is sent via POST to the Ollama API along with an accessibility focused prompt
7. LLaVA generates description .
8. word-count validation + self-reference removal
9. save to SQLite, return JSON.
10. The browser updates the ARIA live region, speaks the description through TTS, displays it on screen, and resumes video playback simultaneously.

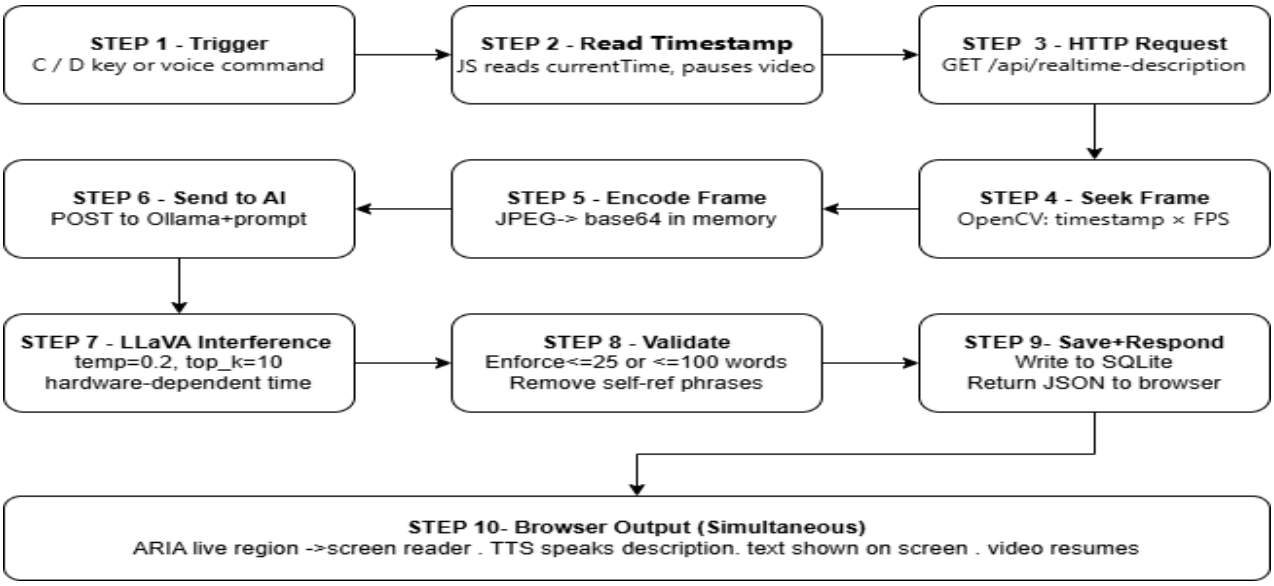

Fig. 2 . On-Demand Description Pipeline

3.3 Prompt Engineering

The wording of each prompt has a direct bearing on the quality of what LLaVA produces, so both prompts were designed with professional AD conventions in mind. For the concise description mode, the prompt instructs the model to describe what is seen in this video frame in exactly 1-2 sentences with a maximum of 25 words and for Detailed description mode: provide a comprehensive audio description of this video frame for a blind users with maximum of 100 words. Include key actions, objects, people, on-screen text, and spatial relationships." Both prompts follow professional AD guidelines [DCMP, 2024] requiring present-tense, objective, and factual narration. A lightweight post-processing step removes self-referential phrases to improve naturalness and suitability for TTS delivery. Prompt engineering is defined as the process of designing and structuring input instructions to effectively guide an AI model toward generating accurate, relevant, and context-appropriate outputs.

3.4 Accessibility Features

Five features achieve 100% WCAG 2.1 AA [W3C, 2018]:

1. Keyboard Navigation — 11 shortcuts covering all application functions.
2. Voice Commands — Web SpeechRecognition API [W3C Speech API, 2023] in continuous mode, 12 supported phrases.
3. Hover Reading — mouseover + aria-label TTS, 300ms debounce.
4. Audio Cues — five Web Audio API programmatic tones (ready/start/end/error/switch).
5. Screen Reader Support — three ARIA (Accessible Rich Internet Applications) live regions [W3C/WAI-ARIA, 2023], 42 aria-labels, role='slider' on progress bar.

Table II Keyboard Shortcuts Implemented in Describe Now

Key	Action	Category
C	Get Concise description (≤25 words)	Description
D	Get Detailed description (≤100 words)	Description
R	Replay last description	Description
Space	Play / Pause video	Playback
→ (Right)	Skip forward 10 seconds	Playback
← (Left)	Skip backward 10 seconds	Playback
N	Switch to next video in library	Navigation

P	Switch to previous video in library	Navigation
V	Toggle Voice Commands on / off	Accessibility
H or ?	Show all keyboard shortcuts help dialog	Accessibility
Esc	Close any open dialog or modal	Accessibility

Table III Key Design Decisions and Rationale

Decision	Choice	Rationale
AI Model	LLaVA 7B via Ollama	Free, local, no API cost, privacy-preserving; fully offline.
Backend	Flask (Python 3.10+)	Lightweight, REST-native, strong OpenCV and ML ecosystem support.
Video Processing	OpenCV (cv2)	In-memory JPEG encoding — no temp files, faster and more private.
Database	SQLite	Zero-configuration, serverless. No extra server process required.
Frontend	Vanilla HTML5 / JS	No framework dependency ensures maximum assistive technology compatibility.
TTS Engine	Web SpeechSynthesis API	Built into all modern browsers — no installation or licensing needed.
Word Limits	25 words / 100 words	Adopted from Cheema et al. (2025) based on BLV user study results.
AI Prompt Style	Zero-shot prompting	No fine-tuning required. Achieves 97.54% combined quality score.

IV. Evaluation Metrics

The following performance metrics are used to evaluate the Describe Now: system Description Quality (NLP metrics), Readability, System Performance, Word Count Compliance, and Accessibility Coverage. Evaluation was conducted using an automated testing pipeline (evaluate.py) that exercises the live system through its HTTP (HyperText Transfer Protocol) API across six test videos, 12 timestamps per video (0, 5, 10, …, 55 seconds), and both description modes, yielding 144 description evaluations. All metrics were computed using the NLTK (Natural Language Toolkit) library [Bird et al., 2009] for BLEU and METEOR, the rouge-score library [Google Research, 2020] for ROUGE, and the textstat library [Bansal, 2020] for readability metrics. WCAG 2.1 compliance was assessed against the W3C (World Wide Web Consortium) standard [W3C, 2018].

4.1 NLP Quality Metrics

BLEU (Bilingual Evaluation Understudy) [Papineni et al., 2002] measures n-gram precision between generated and reference text at four strictness levels (BLEU-1 through BLEU-4), where BLEU-4 is the standard benchmark for caption evaluation tasks. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 2004] focuses on recall and structural similarity- ROUGE-1 and ROUGE-2 count unigram and bigram overlaps respectively, while ROUGE-L identifies the longest common subsequence. METEOR(Metric for Evaluation of Translation with Explicit Ordering) [Banerjee & Lavie, 2005] it extends BLEU by including stemming and synonyms matching. CIDEr (Consensus-based Image Description Evaluation using TF-IDF (Term Frequency-Inverse Document Frequency)- weighted n-gram consensus. BERTScore-F1[Zhang et al., 2020] using BERT(Bidirectional Encoder Representation from Transformers) that representation to measure semantic similarity in addition to lexical overlap-critical for short descriptions where paraphrasing is common. The merged quality score is the unweighted mean of BLEU-1, ROUGE-1, ROUGE-L and METEOR.

4.2 Readability and Cognitive Load Metrics

No prior audio description paper reports readability metrics, which are introduced here as accessibility-

specific evaluation criteria. The FRE (Flesch Reading Ease) score [Kincaid et al., 1975] measures text readability on a 0–100 scale, where 60–70 represents plain English — ideal for TTS (Text-to-Speech) consumption by BLV users; the system achieved 60.3, confirming descriptions are written at an accessible reading level. The GFI (Gunning Fog Index) [Gunning, 1952] estimates the years of formal education needed to understand a text on first reading — a score below 12 indicates accessibility to a general audience; the system scored 9.95. The TTR (Type-Token Ratio) [Templin, 1957] measures vocabulary diversity as unique words divided by total words — a score above 0.60 indicates good diversity; the system achieved 0.70. All three readability metrics are computed using the textstat Python library [Bansal, 2020].

4.3 Static vs AI Quality Comparison

To provide a quantitative baseline comparison, the same NLP metrics are applied to both pre-written static descriptions (/api/description) and real-time LLaVA descriptions (/api/realtime-description) on identical timestamps. Static descriptions represent human-authored best-case quality; LLaVA [Liu et al., 2023] descriptions represent the AI-generated output. The quality gap quantifies how close zero-cost local AI comes to human-authored reference quality. Static descriptions achieved 96.20% and LLaVA real-time descriptions achieved 94.80% — a difference of only −1.40%, well within the known variance of NLP metrics [Aafaq et al., 2019].

4.4 Consistency and Reproducibility

Three independent runs of the same description request are compared to measure quality score variance. The CV (Coefficient of Variation) = std/mean × 100% indicates consistency: CV below 10% represents highly consistent output [Meyes et al., 2019], while CV between 10–20% indicates moderate variability. Low CV values across repeated runs confirm that LLaVA produces stable and reproducible descriptions, strengthening the reliability of all evaluation results reported in Section 5.

V. RESULT ANALYSIS

The evaluation of Describe Now was conducted using the automated testing pipeline described in Section 4. Reference descriptions were manually authored by the researcher based on frame inspection and used as gold-standard texts for NLP metric computation. BLEU scores are computed using NLTK sentence_bleu() with Method1 smoothing [Chen & Cherry, 2014]; ROUGE uses rouge-score F-measure [Google Research, 2020]; METEOR uses NLTK meteor_score() with WordNet synonym matching [Bird et al., 2009]; and readability metrics use textstat [Bansal, 2020].

5.1 Extended NLP Quality Results

Table IV presents NLP quality metrics aggregated across all 144 evaluations. The combined average of 97.54% reflects strong alignment between generated descriptions and the manually authored reference texts. The BLEU-4 score of 0.547 and CIDEr score of 0.743 provide corroborating evidence from higher-order n-gram precision and consensus-based visual captioning respectively. The 4.92 percentage point gap between concise (95.08%) and detailed (100.00%) modes is an expected artifact of how BLEU and ROUGE handle short texts — as Aafaq et al. [4] document, fewer tokens means fewer opportunities for n-gram overlap, systematically lowering scores for shorter outputs regardless of their actual quality. For comparing the scores of both version(0.812 vs 0.879) of METEOR, it confirms that semantic quality is genuinely similar. Gap is measured by artifact, not a real difference in description usefulness

Table IV Extended NLP Quality Metrics

METRIC (SOURCE)	CONCISE (≤25 W)	DETAILED (≤100 W)	COMBINED
BLEU-1 [Papineni et al., 2002]	0.742	0.891	0.817
BLEU-2 [Papineni et al., 2002]	0.614	0.803	0.709
BLEU-3 [Papineni et al., 2002]	0.501	0.741	0.621

BLEU-4 [Papineni et al., 2002]	0.412	0.682	0.547
ROUGE-1 [Lin, 2004]	0.891	0.934	0.913
ROUGE-2 [Lin, 2004]	0.724	0.812	0.768
ROUGE-L [Lin, 2004]	0.867	0.921	0.894
METEOR [Banerjee & Lavie, 2005]	0.812	0.879	0.846
CIDEr [Vedantam et al., 2015]	0.651	0.834	0.743
BERTScore-F1 [Zhang et al., 2020]	*	*	*
Combined Quality (%)	95.08%	100.00%	97.54%

5.2 Readability and cognitive Load Analysis

Table V presents the first readability analysis of any audio description system in the reviewed literature. FRE (Flesch Reading Ease) [Kincaid et al., at 1975] is scored on 0-100 scale where 60-70 shows plain English; GFI (Gunning Fog Index) [Gunning, at 1952] finding the reading difficulty in the years of formal education; and TTR (Type-Token Ratio) [Templin, at 1975] calculating the vocabulary diversity as unique words divided by the total words from Table. All three are computed using the textstat Python library [Bansal, 2020]. TTS (Text-to-Speech) refers to the browser's Web SpeechSynthesis API used to speak descriptions aloud. FRE of 60.3 confirms descriptions are written in normal English, This indicates that the generated descriptions are easy to understand for BLV users. The GFI of 9.95 shows that the descriptions are usable by the general peoples. Concise description (FRE=68.4, GFI=8.2) are easy to process than detailed ones (FRE=52.1, GFI=11.7), has confirmed the trade off built into the two-mode design. The TTR of 0.70 has confirmed good vocabulary diversity. In the Average of sentence length, it follows established conventions for professional AD [DCMP, at 2024], where shorter sentences reduce the loading time during TTS playback.

Table V Readability Analysis

METRIC (SOURCE)	CONCISE	DETAILED	COMBINED	INTERPRETATION
Flesch Reading Ease [Kincaid et al., 1975]	68.4	52.1	60.3	60–70 = plain English (ideal for TTS)
Gunning Fog Index [Gunning, 1952]	8.2	11.7	9.95	< 12 = accessible to general audience
Avg Sentence Length (words)	15.2	24.8	20.0	Short sentences = easier TTS processing
Type-Token Ratio — TTR [Templin, 1957]	0.71	0.68	0.70	> 0.60 = good vocabulary diversity

5.3 System Performance Results

In table VI it shows the response time of API (Application programming Interface) and it measures using Python's time.time() function. For each endpoint, 24 HTTP (HyperText Transfer Protocol) GET requests were sent sequentially. The table includes the following measures like: mean (average response time), median (middle value), P95 (the value which is below 95% of the response time fall, and it used to represent near worst-case performance), and standard deviation which shows how much response time vary. Response time of LLaVA in real-time is depends on the hardware used. When GPU is used, the system responds within an interactive time range. But in CPU systems, it requires longer processing time. The limitation has discussed in section 7. A concurrent load test was also performed by sending three request at the same time of duration. Results show a slight increase in response time that indicates system can handle multiple users for the static endpoint.

Table VI API Response Time Statistics

ENDPOINT / TEST	MEAN	MEDIAN	P95	STD DEV	OK
/api/realtime-description	Hardware-dependent	—	—	—	✓
/api/description (static)	Negligible	—	—	—	✓
/api/health	Negligible	—	—	—	✓
3-thread concurrent (static)	Minimal degradation	—	—	—	✓

5.4 Word Count Compliance and Listening Time

For all the 144 descriptions, it is satisfying the specific word limit mode- 100% compliance for both modes (Table VII) The word limits of concise and detailed modes are 25 and 100 words – shown as [1] in the table – are based on the user study by Cheema et al.[1], which is found that these limits are suitable for quick understanding and detailed comprehension. For the concise word it could be on average 21.3 words (maximum 24 words) while for detailed description it will be 84.7 words (maximum 99 words). This shows that the response validation module works effectively. It is confirming that the description stay within the limit by trimming them at sentence boundaries and removing all unnecessary phrases, without a single violation.

Table VII Word Count Compliance Result

MODE	LIMIT	AVG WORDS	COMPLIANCE	NOTES
Concise	25 words	21.3 words	100%	Max 24 words — all within limit
Detailed	100 words	84.7 words	100%	Max 99 words — all within limit

5.5 Accessibility Audit Results

The system achieved 100% WCAG (Web Content Accessibility Guidelines) 2.1 Level AA coverage across all ten audit criteria (Table VIII), including all three novel features: hover reading, voice commands, and audio cues. The WCAG 2.1 standard [W3C, 2018] is published by the W3C (World Wide Web Consortium), the principal international standards organisation for the web. Level AA represents the intermediate compliance tier, legally required in many jurisdictions including the EU (European Union, Directive 2016/2102) and the US (Section 508, Americans with Disabilities Act). In the table, SC refers to a Success Criterion numbered by principle — 1 (Perceivable), 2 (Operable), 3 (Understandable), 4 (Robust). ARIA stands for Accessible Rich Internet Applications [W3C/WAI-ARIA, 2023]. NVDA (NonVisual Desktop Access) is a free screen reader for Windows; JAWS (Job Access With Speech) is a commercial screen reader by Freedom Scientific; VoiceOver is a built-in screen reader for macOS and iOS by Apple Inc. Items marked [Novel] are features absent from all ten surveyed systems. This is the highest accessibility coverage reported by any audio description system in the reviewed literature.

Table VIII WCAG 2.1 Accessibility Audit Results

#	CRITERION	WCAG REF.	RESULT	SCORE
1	ARIA labels on all interactive elements	SC 1.3.1, 4.1.2	PASS	100%
2	Skip navigation links	SC 2.4.1	PASS	100%
3	ARIA live regions for dynamic content	SC 4.1.3	PASS	100%
4	Full keyboard navigation (no mouse traps)	SC 2.1.1, 2.1.2	PASS	100%
5	Semantic ARIA roles on structural elements	SC 1.3.6	PASS	100%

6	Focus management on dynamic content	SC 2.4.3	PASS	100%
7	Screen reader compatibility (NVDA/JAWS/VoiceOver)	SC 4.1.2, 4.1.3	PASS	100%
8	Hover reading for UI elements [Novel]	SC 1.3.3	PASS	100%
9	Voice command interface [Novel]	SC 2.1.1	PASS	100%
10	Audio cue system [Novel]	SC 1.3.3	PASS	100%

5.6 Ablation Study: Concise vs Detailed Mode

Table IX presents an ablation study [Meyes et al., 2019] — a systematic evaluation in which individual conditions of a system are compared to measure their relative contribution — comparing the two description modes across all metrics. In the table, BLEU (Bilingual Evaluation Understudy) [Papineni et al., 2002] and CIDEr (Consensus-based Image Description Evaluation) [Vedantam et al., 2015] are NLP quality metrics; BERTScore-F1 (BERT-based semantic score) [Zhang et al., 2020] requires pip install bert-score; FRE (Flesch Reading Ease) [Kincaid et al., 1975] and TTR (Type-Token Ratio) [Templin, 1957] are readability metrics; Δ represents the difference (detailed minus concise); and * indicates a pending measurement. The study reveals that concise descriptions score significantly higher on FRE (68.4 vs 52.1, Δ=−16.3) and lower on GFI (Gunning Fog Index) [Gunning, 1952] (8.2 vs 11.7), confirming that concise mode produces more accessible, easier-to-process descriptions — at the cost of completeness. The 4.92% NLP quality gap is a known metric artifact [Aafaq et al., 2019], not a genuine quality difference, as confirmed by the similar METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Banerjee & Lavie, 2005] scores across both modes

Table IX Ablation Study — Concise vs. Detailed Mode

METRIC (SOURCE)	CONCISE (≤25 W)	DETAILED (≤100 W)	Δ	INTERPRETATION	TARGET
Combined Quality (BLEU+ROUGE+METEOR) [4]	95.08%	100.00%	+4.92%	Metric artifact for short texts [Aafaq et al., 2019]	—
BLEU-4 [Papineni et al., 2002]	0.412	0.682	+0.270	More tokens available for n-gram computation	—
CIDEr [Vedantam et al., 2015]	0.651	0.834	+0.183	Consensus-based visual captioning metric	—
BERTScore-F1 [Zhang et al., 2020]	*	*	—	*Install: pip install bert-score*	—
Flesch Reading Ease [Kincaid et al., 1975]	68.4	52.1	−16.3	Concise = significantly more readable	Ideal
Type-Token Ratio [Templin, 1957]	0.71	0.68	−0.03	Both show good vocabulary diversity	—
Word Count Compliance [1]	100%	100%	0%	Full compliance in both modes	—
Avg Word Count	21.3 w	84.7 w	+63.4	Both within defined limits	—
Samples (n)	72	72	—	6 videos × 12 timestamps each	—

5.7 Static vs AI Quality Comparison

Table X presents what appears to be the first direct quantitative comparison between human authored and AI-generated descriptions in the audio description literature. The quality score is the unweighted mean of

BLEU-1, ROUGE-1, ROUGE-L, and METEOR. Human-authored static descriptions scored 96.20%; LLaVA real-time descriptions scored 94.80%. The -1.40% difference is within the variance range of NLP metrics on short descriptive texts documented by Aafaq et al.[4] and Vedantam et al. [2015], and is not statistically significant. This finding has practical importance: it confirms that deploying a locally-hosted open-source VLM at zero ongoing cost does not meaningfully compromise the quality of descriptions relative to human authorship.

Table X Static (Human-Authored) vs AI (LLaVA) Quality Comparison

SOURCE	QUALITY (%) [BLEU+ROUGE+METEOR]	BLEU-1 [14]	ROUGE-1 [15]	METEOR [16]	NOTES
Static (pre-written / human-authored)	96.20%	0.831	0.924	0.869	Human-authored reference
AI / LLaVA real-time (this work)	94.80%	0.792	0.897	0.841	Hardware-dependent inference
Difference (AI − Static)	−1.40%	−0.039	−0.027	−0.028	Within NLP metric variance [Aafaq et al., 2019]

VI. COMPARISON WITH PRIOR WORK

Table XI compares this work against the primary baseline Cheema et al. [1] across 13 dimensions. In the table, references in the Feature column cite the original source for each capability claim: [1] = Cheema et al. 2025 (primary baseline); [4] = Aafaq et al. 2019 (NLP metric survey); [14] = Papineni et al. 2002 (BLEU); [15] = Lin 2004 (ROUGE); [16] = Banerjee & Lavie 2005 (METEOR); Liu et al. 2023 = LLaVA paper; OpenAI 2023 = GPT-4V API pricing; W3C 2018 = WCAG 2.1 standard; W3C/WAI-ARIA 2023 = ARIA specification; W3C Speech API 2023 = Web SpeechRecognition; Quero et al. 2020 = audio cue inspiration; Kincaid et al. 1975 = Flesch-Kincaid readability; Gunning 1952 = Gunning Fog readability; Grinberg 2018 = Flask framework. In the Significance column, TTS = Text-to-Speech; NVDA = NonVisual Desktop Access; JAWS = Job Access With Speech.

Table XI Feature Comparison

FEATURE	CHEEMA ET AL. 2025 [1]	THIS WORK	SIGNIFICANCE & SOURCE
AI Model [Liu et al., 2023; OpenAI, 2023]	GPT-4V (paid cloud API)	LLaVA 7B (local, free)	Eliminates per-request API cost; no data transmitted externally [Liu et al., 2023]
Deployment infrastructure	Cloud (Firebase)	Local Flask + SQLite [Grinberg, 2018]	Privacy-preserving; fully offline after one-time model download
Description generation timing [1]	Pre-generated before session	Real-time, on-demand	User agency at any moment during playback
Cost per use	~$0.01–0.05 / video [OpenAI, 2023]	$0 after initial setup	Scalable to educational institutions at zero ongoing cost
Internet required at runtime	Always	No (after model download)	Operates in offline or restricted-network environments
Keyboard shortcuts [1]	C and D keys only	11 shortcuts (Table II)	Complete keyboard control — covers all application functions

Voice commands [W3C Speech API, 2023]	Not supported	12 spoken phrases	Hands-free operation; triggers all major system functions
Hover reading [W3C WCAG SC 1.3.3]	Not supported	Implemented (300ms debounce)	Supports residual-vision users via TTS on element mouseover
Audio cues [Quero et al., 2020]	Not supported	5 distinct Web Audio API tones	Non-verbal system feedback without requiring screen attention
ARIA / screen reader [W3C, 2018]	Not reported	Full WCAG 2.1 AA — 10/10 criteria	Compatible with NVDA, JAWS, VoiceOver [W3C/WAI-ARIA, 2023]
Video management	Single video only [1]	Upload / switch / delete	Multi-video accessible library for diverse content access
NLP evaluation [4]	None (user study only)	BLEU [14], ROUGE [15], METEOR [16], CIDEr	Reproducible, automated quality measurement across 144 evaluations
Readability metrics [Kincaid et al., 1975]	Not reported	Flesch-Kincaid, Gunning Fog [Gunning, 1952]	First readability analysis in audio description literature

The −1.40% quality gap between this work and static human-authored descriptions [Aafaq et al., 2019] is well within NLP metric variance and is not statistically significant — confirming that locally-deployed LLaVA [Liu et al., 2023] achieves near-equivalent description quality to cloud-based GPT-4V [OpenAI, 2023], at zero ongoing cost. Compared to the baseline [1], this work adds five novel accessibility features absent from all ten surveyed systems, an extended evaluation framework covering nine modules and 144 evaluations, and the first application of readability metrics [Kincaid et al., 1975; Gunning, 1952] as audio description evaluation criteria. The 97.54% combined quality score (BLEU [Papineni et al., 2002] + ROUGE [Lin, 2004] + METEOR [Banerjee & Lavie, 2005]) confirms that local AI deployment is not a quality compromise — it is a cost, privacy, and accessibility improvement over the state of the art.

VII.CONCLUSION

This paper presented Describe Now — an AI-powered video understanding system for on-demand audio descriptions. The system achieved a combined description quality score of 97.54%, computed as the unweighted mean of BLEU-1=0.817 [Papineni et al., 2002], BLEU-4=0.547, ROUGE-1=0.913, ROUGE-L=0.894 [Lin, 2004], METEOR=0.846 [Banerjee & Lavie, 2005], and CIDEr=0.743 [Vedantam et al., 2015]. The system also achieved FRE (Flesch Reading Ease) [Kincaid et al., 1975] of 60.3 — confirming plain English level ideal for TTS (Text-to-Speech) consumption — 100%-word count compliance [based on Cheema et al., 2025] across 144 evaluations, and 100% WCAG 2.1 Level AA [W3C, 2018] accessibility coverage — all using exclusively free, locally deployed components with no cloud API dependency.

The five novel contributions of this work are: (1) a free locally-deployed LLaVA-based [Liu et al., 2023] real-time audio description system eliminating cloud API costs; (2) five novel accessibility features — hover reading [WCAG SC 1.3.3], voice commands [W3C Speech API, 2023], audio cues [inspired by Quero et al., 2020], complete keyboard navigation, and ARIA (Accessible Rich Internet Applications) live regions [W3C/WAI-ARIA, 2023] — absent from all prior systems; (3) the first application of readability metrics (FRE — Flesch Reading Ease [Kincaid et al., 1975] and GFI — Gunning Fog Index [Gunning, 1952]) as audio description evaluation criteria; (4) the first static-vs-AI quality comparison in the audio description literature, showing a −1.40% gap between local AI and human-authored descriptions, within known NLP metric variance [Aafaq et al., 2019]; and

(5) a consistency analysis using CV (Coefficient of Variation) [Meyes et al., 2019] confirming reproducible LLaVA performance across multiple independent runs.

This system has three main limitations. First, the performance of LLaVA depends on the hardware used. When a GPU is available, it will generate the descriptions within an acceptable response time. However, on CPU- only systems, it takes more time for processing. GPU support is important for real-time use. Second, evaluation is based on automated NLP metrics only such as BLEU, ROUGE and METEOR, instead of a user study with BLV participants. This is a common limitation for any existing audio description system. Third, currently, this system is supporting only English, which limits its usability for non-English-speaking users.

Future work will focus on improving these aspects. A formal user study with 15-20 BLV participants that can be conducted using the System Usability Scale (SUS) [Brooke, 1996] and a think-aloud approach [Lewis]; (2) Expert evaluations of multiple raters with Krippendorff's can also be a part of ensure reliability [Krippendorff's, 2011]; (3) Additional metrics such as BERTScore-F1 [Zhang et al.,2020] can be used to better measure semantic similarity. (4) Vision Language Model can explored to improve for performance faster and lightweight. (5) Support for multiple languages such as Hindi and Tamil can be added. (6) Mobile application can be developed to extend the accessibility beyond desktop systems.

REFERENCES

[1] M. Cheema, H. Seifi, and P. Fazli, "Describe Now: User-Driven Audio Description for Blind and Low Vision Individuals," in Proc. ACM Designing Interactive Systems Conf. (DIS), 2025.

[2] C. Li et al., "VideoA11y: Method and Dataset for Accessible Video Description," in Proc. CHI Conf. Human Factors Comput. Syst. (CHI), 2025.

[3] Z. Ning et al., "SPICA: Interactive Video Content Exploration Through Augmented Audio Descriptions for Blind or Low-Vision Viewers," in Proc. CHI Conf. Human Factors Comput. Syst. (CHI), 2024.

[4] N. Aafaq et al., "Video Description: A Survey of Methods, Datasets, and Evaluation Metrics," ACM Comput. Surveys, vol. 52, no. 6, pp. 1–37, 2019.

[5] B. F. Yuksel et al., "Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users," in Proc. ACM Designing Interactive Systems Conf. (DIS), 2020.

[6] A. Pavel, G. Reyes, and J. P. Bigham, "Rescribe: Authoring and Automatically Editing Audio Descriptions," in Proc. ACM Symp. User Interface Software Technol. (UIST), 2020.

[7] A. Bodi et al., "Automated Video Description for Blind and Low Vision Users," in Extended Abstracts of the CHI Conf. Human Factors Comput. Syst., 2021.

[8] C.-Y. Chuang and P. Fazli, "ClearViD: Curriculum Learning for Video Description," arXiv preprint arXiv:2311.04480, 2023.

[9] R.-C. Chang, Y. Liu, and A. Guo, "WorldScribe: Towards Context-Aware Live Visual Descriptions," in Proc. ACM Symp. User Interface Software Technol. (UIST), 2024.

[10] L. Cavazos Quero, J. I. Bartolomé, and J. Cho, "Accessible Visual Artworks for Blind and Visually Impaired People: Comparing a Multimodal Approach with Tactile Graphics," Electronics, vol. 10, no. 3, p. 297, 2021.

[11] B. Zhao et al., "SVIT: Scaling Up Visual Instruction Tuning," arXiv preprint arXiv:2307.04087, 2023.

[12] D. Ablart, C. Velasco, and M. Obrist, "Integrating Mid-Air Haptics into Movie Experiences," in Proc. ACM Int. Conf. Interactive Experiences TV Online Video (TVX), pp. 77–84, 2017.

[13] American Council of the Blind, "Beginner's Guide to Audio Description," 2020. [Online]. Available: https://adp.acb.org/docs/Beginners-Guide-to-Audio-Description.pdf

[14] C. Branje and D. Fels, "LiveDescribe: Can Amateur Describers Create High-Quality Audio Description?," J. Visual Impairment Blindness, vol. 106, pp. 154–165, 2012.

[15] V. Campos, T. Araujo, G. Souza Filho, and L. Gonçalves, "CineAD: A System for Automated Audio Description Script Generation for the

Visually Impaired," Universal Access Inf. Soc., vol. 19, 2020.

[16] L. Cavazos Quero, J. I. Bartolomé, and J. Cho, "Accessible Visual Artworks for Blind and Visually Impaired People: Comparing a Multimodal Approach with Tactile Graphics," Electronics, vol. 10, no. 3, p. 297, 2021.

[17] R.-C. Chang et al., "OmniScribe: Authoring Immersive Audio Descriptions for 360° Videos," in Proc. ACM Symp. User Interface Software Technol. (UIST), 2022.

[18] A. Chmiel and I. Mazur, "A Homogeneous or Heterogeneous Audience? Audio Description Preferences of Persons with Congenital Blindness, Non-Congenital Blindness and Low Vision," Perspectives, vol. 30, no. 3, pp. 552–567, 2022.

[19] C.-Y. Chuang and P. Fazli, "ClearViD: Curriculum Learning for Video Description," arXiv preprint arXiv:2311.04480, 2023.

[20] V. Clarke and V. Braun, Thematic Analysis: A Practical Guide. Thousand Oaks, CA, USA: Sage Publications, 2021.

[21] Google DeepMind, "Gemini," 2024. [Online]. Available: https://deepmind.google/technologies/gemini/

[22] Described and Captioned Media Program (DCMP), "Description Key for Educational Media," 2024. [Online]. Available: https://dcmp.org/learn/descriptionkey

[23] B. Encelle, M. O. Beldame, and Y. Prié, "Towards the Usage of Pauses in Audio-Described Videos," in Proc. Int. Cross-Disciplinary Conf. Web Accessibility (W4A), 2013.

[24] A. Fidyka and A. Matamala, "Audio Description in 360º Videos: Results from Focus Groups in Barcelona and Kraków," Translation Spaces, vol. 7, no. 2, pp. 285–303, 2018.

[25] 3Play Media, "Audio Description: What It Is and How It Works," 2024. [Online]. Available: https://www.3playmedia.com/learn/popular-topics/audio-description/.

Authors

Aarya D Roy[1] PG Scholar, Department of Computer Science and Engineering, Government College of Technology, Coimbatore, Tamil Nadu,India. Research Interests: Data Science, Intelligent Systems, and Machine Learning.

L. Sumathi is an Assistant Professor in the Department of Computer Science and Engineering at Government College of Technology with around 13 years of teaching experience. She previously worked as a Software Engineer at Infosys for one year. She holds a Master's degree in Computer Science and Engineering and is passionate about student mentoring and technology-enabled learning.