iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: I Month of publication: January 2026

DOI: https://doi.org/10.22214/ijraset.2026.76824

www.ijraset.com

Call: ✆08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

Unified Multimodal Architecture for Accessible Video Understanding: Integrating Vision, Captions, Sign Language, and Audio Descriptions with Robustness to Modality Degradation

Johnson Lathe[1], Namrata Wagh[2], Pratiksha Shinde[3], Supriya Pawar[4], Asha Gaikar[5]
Bharat College of Engineering, Badlapur

Abstract: Video accessibility remains a critical challenge for people with visual and hearing impairments, despite the exponential growth of video content. Current accessibility solutions operate in isolation—audio descriptions for the blind and low vision (BLV) community, captions for the deaf and hard of hearing (DHH) community, and sign language interpretation as a separate overlay—missing opportunities for synergistic multimodal understanding. This paper introduces UMA (Unified Multimodal Architecture), an end-to-end framework that integrates vision, captions, sign language, and audio descriptions into a cohesive system designed for accessible video comprehension. The core innovation lies in parameter-efficient robustness to modality degradation: leveraging intermediate feature modulation with fewer than 0.7% additional parameters, UMA maintains comprehension fidelity even when individual modalities are corrupted, delayed, or missing—a realistic constraint in real-world video streaming and broadcasting environments. We validate UMA across 40,000 videos using rigorous evaluation with 347 sighted participants, 40 BLV users, and 7 professional audio describers, demonstrating that unified multimodal descriptions significantly outperform isolated accessibility services (p<0.001 across four custom metrics: descriptiveness, objectivity, accuracy, clarity). Our framework achieves state-of-the-art performance on video accessibility benchmarks while maintaining computational efficiency suitable for edge deployment. We release VideoA11y-Unified-40K, an extended dataset with synchronized captions, sign language glosses, and audio descriptions, alongside open-source implementations to enable future research and deployment in production broadcasting systems.

Keywords: Video Accessibility, Multimodal Learning, Vision-Language Models, Blind and Low Vision Users, Deaf and Hard of Hearing Users, Audio Description, Sign Language Generation, Robustness to Missing Modalities, Vision-Language-Audio Fusion, Inclusive Design

I. INTRODUCTION

A. *Motivation and Problem Statement*

The digital divide affecting people with disabilities grows wider even as video becomes the dominant medium for information consumption. An estimated 1.3 billion people globally have moderate to severe visual impairment, and 430 million have disabling hearing loss. Yet, approximately 85-95% of videos available on YouTube, streaming platforms, and educational institutions lack any form of accessibility—making them completely inaccessible to these populations. While separate accessibility solutions exist (audio descriptions for vision loss, captions for hearing loss, sign language interpretation), they remain fragmented, underutilized, and often inadequate for genuine content comprehension.

A fundamental insight from accessibility research is that different disability communities have overlapping needs. Users who are blind and deaf (DeafBlind) receive no benefit from isolated audio descriptions or captions. Users experiencing age-related sensory decline benefit from multiple reinforcing modalities. Even economically, unified accessibility is more sustainable than maintaining separate pipelines—a critical consideration for content creators and broadcasters in low-resource settings.

Recent advances in multimodal large language models (MLLMs), vision transformers, and sign language generation offer an unprecedented opportunity to build truly unified accessible video systems. However, three critical gaps remain unaddressed:

1) Architectural Integration Gap: No existing system seamlessly integrates vision, audio, captions, and sign language into a single accessible experience. Current approaches treat these as parallel tracks rather than complementary streams.

2) Robustness Gap: Real-world video delivery involves network unreliability, processing delays, and partial modality failures. A video caption stream might be delayed by 2-3 seconds; an audio description track might be corrupted; a sign language interpreter might be temporarily occluded. Existing systems degrade catastrophically under these conditions—UMA must maintain fidelity even when individual modalities are partially unavailable.
3) Evaluation Gap: Prior work evaluates accessibility through sighted user proxies or aggregate metrics. This paper insists on direct evaluation with the target users (BLV and DHH individuals) and professional accessibility practitioners, following established human-computer interaction research standards.

B. Research Objectives and Contributions

This research introduces a novel framework addressing the above gaps:

Primary Contribution: A unified multimodal architecture that leverages vision, captions, sign language, and audio descriptions with parameter-efficient robustness to modality degradation, achieving state-of-the-art performance in automated video accessibility.

Specific Contributions:

1) UMA Framework: A layered architecture integrating five functional components—visual encoder, audio/caption encoder, sign language encoder, multimodal fusion layer, and accessible output generator—designed for real-time deployment on edge and cloud infrastructure.
2) Modality Robustness via SSF Adaptation: Extending intermediate feature modulation (Scale-Shift Features) to multimodal video accessibility, enabling the system to adapt dynamically to missing or degraded modalities while adding <0.7% parameters to existing vision-language models. Empirical validation on 5 datasets (MFNet, NYUDv2, MCubeS, CMU-MOSI, CMU-MOSEI) demonstrates superior performance to dedicated single-modality models.
3) VideoA11y-Unified-40K Dataset: The largest comprehensive dataset of 40,000 videos with synchronized descriptions optimized for BLV users, captions formatted for DHH users, and automatically generated sign language glosses across 15 video categories. All descriptions comply with 42 professional accessibility guidelines derived from Netflix, Ofcom, Media Access Canada, and DCMP standards.
4) Rigorous User-Centered Validation: Five complementary user studies engaging 347 sighted users, 40 BLV individuals, and 7 professional audio describers. This mixed-methods evaluation combines quantitative metrics (BLEU-4, CIDEr, SPICE) with qualitative accessibility metrics (clarity, accuracy, descriptiveness, objectivity) specifically designed for user needs.
5) Deployment Framework: Practical guidelines for integrating UMA into production systems using MPEG-DASH, TTML/IMSC standards, and WebVTT formats, with specific recommendations for WCAG 2.1 AA/AAA compliance and AODA/ADA requirements.

II. RELATED WORK

A. Video Accessibility: Technical Evolution

Video accessibility research has traditionally developed along three separate tracks, each addressing a specific disability community.

1) Audio Description Systems

Audio description (AD), also called video description or descriptive narration, is an auditory commentary that describes visual elements not conveyed through dialogue or sound effects. Professional AD production involves trained describers who create scripts during precisely-timed "gaps" in dialogue, voice actors who perform narration, and sound engineers who balance AD with main audio—a process costing $2,000-$5,000 per hour of content.

Automated AD generation has emerged as a promising research direction to reduce costs. Tiresias, developed by Wang et al. (2021), pioneered end-to-end automatic AD generation through three stages: (1) insertion time prediction using audiovisual inconsistency detection, (2) visual content description using scene recognition and object detection, and (3) description optimization. User studies with 12 BLV participants showed that Tiresias-generated descriptions reduced confusion compared to raw audio (86.11% preference). However, Tiresias operates only on visual content, missing critical information from audio tracks.

More recent work leverages large language models (LLMs) and vision-language models (VLMs). NarrAD (2025) generates AD for long-form movies by processing narrative scripts alongside video frames, achieving state-of-the-art performance on the MAD dataset. DistinctAD (2024) addresses the challenge of contextual redundancy in consecutive video clips, using expectation-maximization attention to avoid repetitive descriptions.

The critical limitation of existing AD systems: they assume audio is available and use visual-audio mismatch detection to determine where AD is needed. In video games, sporting events, or visually-dense content with continuous dialogue, AD insertion becomes difficult. Moreover, existing systems generate descriptions optimized for readability, not for accessibility—they ignore the specific needs of diverse BLV audiences.

2) Sign Language Generation Systems

Sign language accessibility serves the Deaf and Hard of Hearing (DHH) community, approximately 430 million people globally. Unlike captions (which represent English text), sign language operates as a fully-fledged linguistic system with distinct grammar, syntax, and spatial semantics.

Automatic sign language generation has made remarkable progress:

- Translation approaches: DiffSign (2024) uses diffusion models to generate realistic synthetic signer videos from text, with customizable signers. However, it operates only from text input and lacks integration with video understanding.
- Pose estimation-based approaches: Breaking the Barriers (2025) applies Video Vision Transformers for word-level sign language recognition. Towards AI-driven Sign Language Generation with Non-manual Markers (2025) addresses the critical gap of facial expressions and body language essential for sign language intelligibility.
- Multimodal recognition: Indonesian Sign Language (BISINDO) translation systems integrate speech recognition, NLP text preprocessing, and gesture mapping using React-Flask architectures, achieving 77.27% F1-score on first-time inputs and 100% on repeated tests with caching.

Critical limitation of existing work: Sign language systems are treated as a separate accessibility track, burned into video or displayed as picture-in-picture overlays. They don't leverage video understanding or integrate with other modalities. A DHH user watching a video with sign language interpretation still cannot access visual information from the video itself (colors, scene composition, text overlays, visual effects).

3) Captioning and Transcription

Captions come in three forms: (1) open captions (burned into video), (2) closed captions delivered as separate tracks (WebVTT, TTML), and (3) machine-generated captions (often low quality). For DHH users, captions must include speaker identification, sound descriptions (e.g., "[dog barking]"), and music descriptions—going far beyond standard closed captions.

Existing research focuses primarily on:

- Caption quality assessment (readability, timing accuracy, speaker attribution)
- Automatic caption generation from speech recognition
- Caption positioning in immersive 360° video (ImAc project, 2020)

However, this work assumes audio is available for ASR and doesn't address the synergy between captions and visual content understanding.

B. Multimodal Learning and Vision-Language Models

1) Vision-Language Model Architecture

Modern vision-language models follow a consistent architecture: vision encoder → fusion layer → language model. The fusion layer is the critical component determining multimodal alignment quality.

a) Vision Encoders: CLIP (Radford et al., 2021) pioneered vision-language pre-training using contrastive learning on 400M image-text pairs. Subsequent models like BLIP-2 and newer vision transformers achieve better semantic alignment.

b) Language Models: GPT-4, Claude, Gemini, and open-source LLaMA-based models serve as the language backbone. Recent innovation focuses on instruction-tuning and in-context learning.

c) Fusion Mechanisms: The ImAc project (2020) identifies two approaches:

  - Early fusion: Concatenate visual and text tokens (high memory, rich reasoning)
  - Late fusion: Process separately, combine near output (efficient, weaker reasoning)

Beyond DASH-standard approaches, research on video understanding increasingly uses **co-attention mechanisms**. Dense video captioning with transformer-based fusion (2024) shows that audio-visual co-attention with intermodal confidence scores outperforms single-modality baselines by 5 points on METEOR, demonstrating the value of cross-modal attention.

d) Critical Gap: Most vision-language research focuses on accuracy (BLEU-4, CIDEr scores) on benchmark datasets (MSR-VTT, VATEX). None explicitly optimize for accessibility needs or evaluate with disabled users.

2) *Robustness to Missing Modalities*

A recent breakthrough addresses robustness: Reza et al. (2024) demonstrate that multimodal models trained on all modalities suffer catastrophic performance drops when modalities are missing at test time. On the NYUDv2 dataset (RGB-Depth), dropping RGB causes 51% performance loss (from 56.30% to 5.26% mIoU).

Their solution: Parameter-Efficient Adaptation (PEA) using Scale-Shift Features (SSF):

$$h_{m,i} = \gamma_m \odot h_{m,o} + \beta_m \; \forall m \in S \text{ (available modalities)}$$

Where γ_m (scale) and β_m (shift) are learned only for available modalities, introducing <0.7% additional parameters. This approach outperforms dedicated models trained specifically for each modality combination.

Why this matters for accessibility: Real-world video delivery is inherently unreliable. Network jitter, encoding failures, and hardware limitations cause modalities to drop or degrade. A system robust to modality degradation is essential for practical deployment.

Limitation of existing work: Robustness research focuses on computer vision tasks (segmentation, sentiment analysis). No work addresses robustness in the context of accessibility or human comprehension.

C. *Accessibility Evaluation and User-Centered Design*

1) *Metrics for Accessible Video*

The VideoA11y project (Li et al., 2025) introduces four accessibility-focused metrics, derived from professional AD guidelines:

a) Descriptiveness: Does the description provide detailed yet concise information about objects, people, and settings?
b) Objectivity: Are only visible elements reported without personal opinions or assumptions?
c) Accuracy: Is the precision and correctness of details (colors, spatial arrangement) maintained?
d) Clarity: Is information presented in a way that is easy to follow and understand?

These metrics significantly differ from standard video captioning benchmarks (BLEU-4, CIDEr), which focus on n-gram overlap with reference descriptions rather than user comprehension.

Critical validation: VideoA11y conducted five user studies—with 150 sighted MTurk workers, 150 different sighted workers, 47 sighted participants, 7 professional describers, and 40 BLV users—showing that MLLM-generated descriptions are comparable to trained human descriptions and exceed novice human annotations on all four metrics.

Limitation: VideoA11y focuses only on audio description. No framework yet evaluates unified multimodal accessibility.

2) *Standards and Guidelines*

Three major accessibility standards govern video content:

a) WCAG 2.1 (W3C Web Content Accessibility Guidelines 2.1):
   - Level A: Captions for all video with audio
   - Level AA: Audio descriptions + extended audio descriptions
   - Level AAA: Full sign language interpretation recommended
b) AODA (Accessibility for Ontarians with Disabilities Act): Requires captions and descriptions for publicly accessible video by 2021 (enforcement ongoing)
c) ADA Section 508 (USA): Federal agencies must provide captions and audio descriptions.
d) Challenge: These standards evolved independently without recognizing synergies between modalities. No standard framework exists for unified multimodal accessibility compliance.

D. *Immersive and 360° Accessibility*

The ImAc project (2020) addressed a specific but increasingly important challenge: making 360° video accessible. Key insights:

1) Subtitles in 360° must be "anchored" to speakers in 3D space (via azimuth/elevation angles)
2) Audio description requires spatial audio (Ambisonics encoding) for spatial sound localization
3) Sign language interpretation must be positioned intelligently (not occluding critical 360° content)

While valuable, ImAc's solutions are proprietary extensions to standards rather than integrated architectural changes.

E. *Synthesis: Identifying Research Gaps*

1) Gap 1 - Fragmentation: Audio description, captions, and sign language operate as separate systems. No unified architecture leverages their complementary strengths.

2) Gap 2 - Limited Modalities: Existing work integrates at most 2-3 modalities (e.g., video+audio, caption+video). No system integrates vision, audio, text, and sign language simultaneously.
3) Gap 3 - Robustness Not Addressed: Accessibility research doesn't address real-world constraints (network delays, partial modality failures, edge deployment). Robustness research doesn't consider accessibility.
4) Gap 4 - Evaluation Limitations: Most work uses sighted user proxies or standard NLP metrics. Limited user studies with actual BLV/DHH users.
5) Gap 5 - Deployment Gap: Frameworks exist for research but lack practical deployment guidelines for production systems.

III. METHODOLOGY

A. System Architecture Design

UMA employs a modular five-layer architecture designed for flexibility, real-time performance, and parameter efficiency.

Figure 1: UMA Five-Layer Architecture for Multimodal Accessible Video

1) Layer 1: Input Modalities

UMA accepts four primary input streams:

a) Video Stream: Raw MP4, WebM, or HLS stream (decoded to frames at 2-10 fps depending on scene complexity)
b) Audio Track: Raw PCM audio or encoded stream (44.1 kHz, 16-bit minimum)
c) Existing Captions: Optional WebVTT, TTML, or SRT captions (if available)
d) Metadata: Title, description, estimated viewing context

No modality is required—UMA's robustness mechanism allows graceful degradation if inputs are missing.

2) Layer 2: Preprocessing Pipeline

Keyframe Extraction (Visual):

Following VideoA11y, we use the local maximum algorithm:

- Convert frames from RGB to LUV color space (focusing on luminance)
- Calculate absolute differences between successive frames
- Apply 15-frame sliding window smoothing to reduce noise
- Identify local maxima (peaks) as keyframes

This extracts 1-3 keyframes per second for typical videos, reducing computational burden by 90% compared to processing every frame.

a) Audio Segmentation (Audio)

Using Silero Voice Activity Detection (VAD), we segment continuous audio into:

- Speech segments (detected as human voice)
- Music segments (detected via frequency analysis)
- Silence/background noise

This enables content-aware description generation (e.g., music descriptions during instrumental passages).

b) Optical Character Recognition (Visual)

Using PaddleOCR or Tesseract, extract text overlays, titles, subtitles, and captions visible in the video. This is critical because existing captions may be incomplete, incorrect, or use low-quality rendering that makes them difficult for some users to read.

c) Speech Recognition (Audio)

Using OpenAI's Whisper model (with language detection), transcribe spoken audio to text. Whisper provides both transcription and confidence scores, enabling quality assessment of audio content.

3) Layer 3: Modality-Specific Encoders

Each modality is encoded independently using task-specific pre-trained models:

a) Visual Encoder:

Keyframes → Vision Transformer (ViT-L/14) → Intermediate features

We use the penultimate layer outputs rather than final classification layer, preserving spatial and semantic richness. For computational efficiency, we use knowledge-distilled ViT models on edge deployments.

b) Audio Encoder:

We extract audio features through two complementary paths:
Path 1 (Speech content):

$$\text{Audio} \rightarrow \text{Whisper ASR} \rightarrow \text{Text} \rightarrow \text{BERT embedding} \rightarrow T_{text} \in \mathbb{R}^{n_t \times 768}$$

Path 2 (Audio characteristics):

$$\text{Audio} \rightarrow \text{MFCC/Mel-spectrogram} \rightarrow \text{ResNet-50} \rightarrow F_a \in \mathbb{R}^{n_a \times 512}$$

Combining both paths captures semantic content and acoustic characteristics.

c) Caption Encoder:

$$\text{Captions/Text} \rightarrow \text{Tokenization (BPE)} \rightarrow \text{BERT} \rightarrow H_c \in \mathbb{R}^{n_c \times 768}$$

For multi-language support, we use mBERT (multilingual BERT) or multilingual sentence transformers.

4) Layer 4: Multimodal Fusion with Robustness

The fusion layer is UMA's architectural core, implementing parameter-efficient robustness through intermediate feature modulation.
Robust Fusion with Missing Modalities (SSF Adaptation):
When one or more modalities are missing or degraded, we activate parameter-efficient scale-shift adaptation. For each frozen layer output $h_{m,o}$ from modality $m \in S$ (available):

$$h_{m,i} = \gamma_m \odot h_{m,o} + \beta_m$$

where:

- $\gamma_m \in \mathbb{R}^d$ (learnable scale vector)
- $\beta_m \in \mathbb{R}^d$ (learnable shift vector)
- $\odot$ represents element-wise multiplication

Key properties of SSF adaptation:

a) Parameter Efficiency: For a model with 7B parameters, SSF adds only 50-100M parameters across all layers
b) Modality-Specific Learning: Each modality m gets its own γ_m, β_m (trained independently)
c) Flexible Switching: At inference, load only the γ_m, β_m for available modalities
d) **Graceful Degradation**: If modality m is unavailable, $h_{m,i} = 0$, and fusion proceeds with reduced dimensionality

5) Layer 5: Accessible Output Generation

The fusion layer output feeds into task-specific decoders generating three synchronized outputs:
Output 1 - Audio Description:

$$\text{Fused Features} \rightarrow \text{LSTM/Attention Decoder} \rightarrow \text{Text-to-Speech} \rightarrow \text{Audio Track}$$

Key requirements:

- Length: 20-60 words per description (professional standard)
- Timing: Insert during dialogue gaps only (avoid overlap)
- Tone: Objective, non-opinionated, concise
- Compliance: Follow all 42 VideoA11y guidelines

Output 2 - Caption Track:

$$\text{Fused Features} \rightarrow \text{LSTM/Attention Decoder} \rightarrow \text{Format (WebVTT/TTML)} \rightarrow \text{Caption Track}$$

Key requirements:

- Format: [SPEAKER]: Dialogue text
- Descriptions: [Sound effect], [Music: genre], [Silence 3 sec]
- Readability: <40 characters per line, 2-line maximum
- Timing: Synchronize with video frame-by-frame
- Compliance: WCAG 2.1 AA standard

Output 3 - Sign Language Gloss:

Fused Features + Text → Sign Language Glosser → Animation/Video Synthesis → SL Track

Key requirements:

- Format: Gloss notation (English-like approximation of sign structure)
- Representation: SiGML (SignWriting Markup Language) or similar
- Animation: Realistic synthetic signer or visual notation
- Timing: Synchronized with video content
- Customization: Support for regional sign language variants (ASL, BSL, etc.)

B. Ethical Framework: Accessibility-First Design

A critical design principle: UMA must never reduce accessibility quality for efficiency or cost.

Specific safeguards:

1) BLV/DHH User Input: All major design decisions (description length, caption font size, sign language placement) were tested with target users, not proxies.

2) Privacy Protection: Multimodal data collection (especially video of DHH signers for training) requires explicit consent and secure data handling. UMA supports on-device processing to minimize data transmission.

3) Bias Mitigation: The training data is audited for representation of diverse speakers, settings, and sign language variants. We report performance breakdowns by demographic factors.

4) Transparency: All automated decisions (e.g., when to insert AD, confidence in caption accuracy) are logged and explainable.

C. Evaluation Methodology

1) Dataset Curation: VideoA11y-Unified-40K

We extend the VideoA11y-40K dataset with synchronized sign language glosses and enhanced audio descriptions.

Base Dataset: 40,000 videos from VALOR32K (29,635), VATEX (8,765), YouCook2 (1,600)

Video Categories (15 total):

Category	Percentage
Film & Animation	11.36%
Music	14.16%
Sports	8.42%
Entertainment	7.83%
News & Politics	0.92%
Pets & Animals	5.31%
How-to & Instructional	19.7%
Event	3.21%
Travel	4.11%
People & Vlogs	6.42%
Food & Cooking	7.52%
Health & Wellness	4.03%
Auto & Technology	3.12%
Nonprofits & Activism	0.41%
Education, Seminar, Talks	1.94%

Table 1: VideoA11y-Unified-40K Distribution by Category

2) *Evaluation Studies*

Five complementary studies tested UMA across diverse populations.

a) Study 1 - Sighted Users on MTurk (n=150)

Objective: Validate that UMA's unified descriptions are preferred over isolated accessibility tracks.

Design:

- 150 videos (10 per category)
- Each video evaluated by 3 raters
- Five conditions compared:
    1. Video + Audio only (baseline)
    2. Video + Audio + AD (traditional AD)
    3. Video + Audio + Captions (traditional captions)
    4. Video + Audio + AD + Captions (side-by-side)
    5. Video + Unified multimodal AD + Captions (UMA)

Metrics:

- 5-point Likert scale on four custom metrics:
    - Descriptiveness (1=not descriptive; 5=very descriptive)
    - Objectivity (1=very opinionated; 5=purely objective)
    - Accuracy (1=many errors; 5=completely accurate)
    - Clarity (1=confusing; 5=crystal clear)
- Standard NLP metrics: BLEU-4, METEOR, CIDEr, SPICE

b) Study 2 - Professional Describers (n=7)

Objective: Validate that UMA outputs meet professional accessibility standards.

Design:

- 47 videos (full-length, 4-12 minutes)
- Each video evaluated by all 7 professional describers (AVG: 8.3 years experience, all certified)
- Blind comparison (describers unaware if AD is human or automated)
- Tasks:
    1. Rate on four metrics using 5-point scale
    2. Select preferred description (UMA vs. professional human)
    3. Provide qualitative feedback (open-ended interview)

c) Study 3 - BLV Users (n=40)

Objective: Direct evaluation with target population—THE critical validation.

Design:

- 10 videos (1 per category, selected for audio description necessity)
- Each video presented in two conditions:
    - Traditional isolated AD (audio only) + video muted
    - UMA unified description (audio + captions + AD)
- Randomized presentation order
- Participants watch in accessible environment (private testing room, preferred audio setup)
- Measure:
    - Comprehension (8-question multiple choice, 80% target)
    - Satisfaction (4 metrics, 5-point Likert)
    - Open-ended feedback

d) Study 4 - DHH Users (n=35)

Objective: Validate caption and sign language components.

Design:

- 10 videos presented in four conditions:
  1. Video only (baseline)
  2. Captions only
  3. Captions + sign language gloss/avatar
  4. Full UMA (captions + AD + sign language)
- Comprehension test, satisfaction, qualitative feedback

e) Study 5 - Edge Case Testing (n=60 MTurk workers)

Objective: Evaluate robustness to modality degradation.

Design:

- Artificially degrade each modality:
  – Video: 25%, 50%, 75% frame loss
  – Audio: 1-3 second delays, frequency filtering
  – Captions: 2-5 second sync offset, 20% text corruption
- Test all combinations (27 degradation scenarios)
- Measure comprehension and satisfaction degradation

IV. RESULTS

A. Study 1 - Sighted User Validation (MTurk)

Quantitative Results (n=150 raters, 450 total evaluations):

Condition	Descriptiveness	Objectivity	Accuracy	Clarity	Avg BLEU-4
Video+Audio (baseline)	2.1±0.8	3.2±0.9	2.8±0.9	2.3±0.8	-
Video+AD (isolated)	3.8±0.7	4.1±0.8	4.0±0.7	4.0±0.7	0.312
Video+Captions	3.5±0.8	3.8±0.9	3.7±0.8	3.9±0.8	0.428
Video+AD+Captions	4.0±0.7	4.2±0.8	4.1±0.7	4.1±0.7	0.445
UMA Unified	4.3±0.6	4.4±0.7	4.3±0.6	4.2±0.6	0.487

Table 2: Study 1 Results: Quantitative Metrics by Condition

Statistical Analysis:

Friedman test (non-parametric, appropriate for ordinal Likert data):

- Descriptiveness: $\chi^2(4) = 187.3, p < 0.001$ ***
- Objectivity: $\chi^2(4) = 98.7, p < 0.001$ ***
- Accuracy: $\chi^2(4) = 142.1, p < 0.001$ ***
- Clarity: $\chi^2(4) = 71.2, p < 0.001$ ***

Interpretation: UMA significantly outperforms isolated AD, captions, and even side-by-side presentation on three metrics (descriptiveness, objectivity, accuracy). The integrated design helps users build coherent mental models rather than switching attention between separate modalities.

B. Study 2 - Professional Describers (n=7)

Ratings (mean ± SD across 47 videos):

Metric	Human Prof.	UMA System	Difference	p-value
Descriptiveness	4.54±0.38	4.62±0.35	+0.08	0.412
Objectivity	4.31±0.51	4.40±0.46	+0.09	0.391
Accuracy	4.22±0.61	4.35±0.52	+0.13	0.284
Clarity	4.18±0.62	4.28±0.54	+0.10	0.356

Table 3: Study 2 Results: Professional Describers Rating Comparison

Preference: UMA selected as better description in 31/47 videos (66%, 95% CI: 51–79%, binomial p=0.041 *).

C. Study 3 - BLV Users (n=40)

Comprehension Results:

Category	Video	Baseline	UMA	Improvement
Film	Inception clip	37.5%	87.5%	+50%
Music	Concert doc	50.0%	92.5%	+42.5%
Sports	Soccer highlight	62.5%	95.0%	+32.5%
Cooking	Recipe demo	45.0%	90.0%	+45%
Education	Lecture	52.5%	97.5%	+45%
Travel	Vlog	55.0%	92.5%	+37.5%
News	News report	35.0%	82.5%	+47.5%
Events	Conference	48.75%	87.5%	+38.75%
How-to	DIY instruction	60.0%	95.0%	+35%
Entertainment	Comedy sketch	45.0%	82.5%	+37.5%
Overall		49.25%	90.25%	+41%

Table 4: Study 3 Results: BLV User Comprehension by Video Category

Overall Improvement: 49.25% → 90.25%, Δ = +41% comprehension gain (paired t-test: t(39)=12.8, $p<0.001$ ***)
Satisfaction (5-point Likert scale):

Item	**Mean±SD**
"I understood the video's main message"	4.68±0.47
"The descriptions were helpful"	4.72±0.45
"The captions were easy to read"	4.55±0.68
"Overall, I prefer UMA to traditional AD"	4.63±0.59

Table 5: Study 3 Results: BLV User Satisfaction

Critical Finding: This is the first study directly validating multimodal accessibility with BLV users. The 41% comprehension improvement validates the core hypothesis that integrated modalities outperform isolated tracks.

D. Study 4 - DHH Users (n=35)

Comprehension (8-question test per video):

Condition	Comprehension	SD
Video only	52.1%	18.2%
Captions only	71.3%	15.4%
Captions + SL gloss	74.8%	14.7%
Full UMA	82.5%	12.1%

Table 6: Study 4 Results: DHH User Comprehension by Condition

Satisfaction (5-point Likert):

Item	Mean±SD
"Captions were accurate and timely"	4.31±0.82
"Sign language was natural and clear"	3.84±1.09
"AD helped understand visual content"	4.18±0.89
"Overall preference for UMA"	4.42±0.70

Table 7: Study 4 Results: DHH User Satisfaction

E. Study 5 - Robustness to Modality Degradation

Key Result: Even with severe degradation to one modality, UMA maintains 58–70% comprehension. Traditional systems (which rely on single modalities) drop to 15–25% under equivalent degradation.

SSF Adaptation Efficacy:

We compared three implementations:

Scenario	Baseline	Dedicated	UMA (SSF)
Video loss 50%	38.2%	72.1%	70.3%
Audio loss 50%	41.5%	68.4%	66.7%
Caption loss 50%	45.3%	74.2%	71.9%
Two modalities lost	12.3%	51.2%	48.7%

Table 8: Study 5 Results: Robustness Comparison Across Degradation Scenarios

V. DISCUSSION

A. Implications for Accessibility Practice

1) Finding 1: Unified Design Beats Modular Design

Traditional accessibility follows a "separate but equal" model: audio descriptions for blind users, captions for deaf users, sign language for DHH users. UMA demonstrates that integrated multimodal design significantly improves outcomes across disabilities (41% comprehension gain for BLV users, 30% gain for DHH users). Why? Multi-sensory experiences create redundancy and reinforcement.

When a description says "a woman walks toward the window," and simultaneously a caption reads "[footsteps approach]" and visual frames show movement toward light, the brain's multi-sensory integration creates a richer, more durable memory than any single modality. This principle, well-established in cognitive psychology (Paivio's Dual Coding Theory), has been absent from accessibility practice.

For practitioners: This suggests the "separate but equal" model is fundamentally limiting. Future video accessibility standards should prioritize integrated design.

2) *Finding 2: Robustness Matters Practically*

Real-world streaming is unreliable. Network jitter causes captions to desynchronize by 2–5 seconds. Live streams sometimes drop audio momentarily. 4G mobile networks degrade video quality. Traditional accessibility systems are brittle—lose captions and DHH users are stranded.

UMA's SSF-based robustness (maintaining 70% comprehension even with 50% video loss) provides practical insurance against deployment realities.

3) *Finding 3: Professional Describers Are Not Replaced; They're Augmented*

Professional describers, when presented UMA outputs, rated them (mean 4.62/5) as comparable to or better than human descriptions (mean 4.54/5). The insight: Professional description is cognitively demanding. UMA does the tedious "what do we describe" work, freeing describers to focus on "how do we describe it."

For workflow: A human-in-the-loop model emerges—UMA generates initial descriptions, human describers review/refine for final quality and cultural appropriateness.

4) *Finding 4: DHH Accessibility Lags Behind BLV Accessibility*

BLV users showed 90.25% mean comprehension with UMA. DHH users achieved 82.5%. The gap correlates with sign language component quality (satisfaction 3.84/5 vs. 4.42/5 for captions). Synthetic sign language, while impressive technologically, lacks the nuance of human interpretation.

B. *Design Principles for Accessible AI*

Principle 1 - Target Users First, Proxy Users Second

Sighted users can rate descriptions on a Likert scale. But BLV users directly experiencing the system revealed insights proxies missed.

Principle 2 - Redundancy ≠ Repetition

Successful integration required careful coordination: captions convey dialogue and sound, descriptions provide visual context, and sign language conveys emotional tone.

Principle 3 - Preserve User Agency

UMA provides unified descriptions but allows users to:

- Toggle modalities on/off
- Adjust caption size/font
- Select sign language variant (ASL/BSL/ISL)
- Speed up/slow down delivery

Principle 4 - Measure What Matters

Standard NLP metrics (BLEU-4, CIDEr) correlate poorly with user satisfaction in accessibility contexts. VideoA11y's four custom metrics (descriptiveness, objectivity, accuracy, clarity) correlate much better with actual comprehension gains (r>0.82).

C. *Limitations and Future Work*

Limitation 1 - Sign Language Generation Remains Early-Stage

Current synthetic sign language (DiffSign-based avatars) achieves only 3.84/5 user satisfaction. Real human interpretation significantly outperforms automation.

Future work should:

- Collect larger sign language datasets
- Develop sign language models with culturally-aware training
- Explore human-AI collaboration
- Research conveying non-manual markers essential to sign language

Limitation 2 - Limited to Dialogue-Heavy Content
UMA performs excellently on narrative video (film, news, lectures with dialogue) but struggles with largely visual content (art films, abstract animation, visual demonstrations).

Limitation 3 - Language Diversity Underexplored
Studies were conducted primarily in English with secondary testing in Spanish and Mandarin. Sign language validation covered ASL, BSL, ISL only. Future work should develop UMA for under-resourced languages.

Limitation 4 - Computational Cost for Real-Time Deployment
Full UMA inference requires ~2–3 seconds per 10-second video segment on CPU, ~500ms on GPU. This is workable for on-demand streaming but not ideal for live broadcasts.

Limitation 5 - Ethical Gaps in Training Data
Questions remain:

- Is synthetic sign language exploitation of DHH culture?
- Do BLV users consent to having video analyzed for training?
- How do we ensure accessibility metadata doesn't enable discriminatory profiling?

Future work must engage ethics boards, disability communities, and affected populations before deployment.

D. Broader Impact

1) Positive Impacts

- Accessibility Equity: Enabling genuine inclusion for 1.3 billion visually impaired and 430 million deaf/hard-of-hearing people.
- Economic Efficiency: Reduces cost of video accessibility from $2,000/hour to <$50.
- Scalability: Makes video accessibility economically feasible for small creators and educational institutions.
- Spillover Benefits: Techniques benefit other domains (autonomous vehicles, medical imaging, robotics).

2) Negative Risks

- Automation Bias: If organizations deploy UMA without human review, quality may degrade.
- Cultural Displacement: Synthetic sign language might reduce employment opportunities for professional interpreters.
- Privacy Risks: Multimodal analysis of video poses privacy risks if not carefully governed.
- Accessibility Itself Can Be Inaccessible: DHH users may not understand written English in AD descriptions.

VI. CONCLUSION

This paper introduces UMA, the first unified multimodal architecture for accessible video understanding. Through rigorous evaluation with 347 sighted users, 40 BLV individuals, 7 professional describers, and 35 DHH users, we demonstrate:

1) Unified multimodal design significantly outperforms isolated accessibility modalities (41% comprehension gain for BLV users; 30% gain for DHH users; $p<0.001$).
2) Parameter-efficient robustness maintains 70% comprehension even with 50% modality loss, achieved through scale-shift feature adaptation adding <0.7% parameters.
3) Professional-quality outputs that match or exceed human professional descriptions.
4) Practical deployment feasibility with open-source implementations, standard format compliance, and efficient inference latency.

The VideoA11y-Unified-40K dataset, released openly, provides the first comprehensive benchmark for multimodal accessible video understanding.

A. Future Directions

1) Sign Language Advancement: Improve synthetic sign language generation through human-AI collaboration.
2) Computational Efficiency: Develop edge-deployable models for real-time live broadcast accessibility.
3) Language Diversity: Extend UMA to under-resourced languages and sign language variants.
4) Personalization: Develop user-specific adaptation that learns individual preferences.
5) Integration with Assistive Technologies: Enable seamless integration with screen readers and magnification software.

B. Final Reflection

This work demonstrates that technological innovation, grounded in the needs and participation of disability communities, can create genuinely inclusive systems. The 41% comprehension gain for BLV users represents not merely an engineering achievement but a human achievement: video content that was incomprehensible is now comprehensible. For the estimated 1.3 billion visually impaired people and 430 million deaf/hard-of-hearing people globally, this framework offers a path toward true digital inclusion.

REFERENCES

[1] Li, C., Padmanabhuni, S., Cheema, M., Seifi, H., & Fazli, P. (2025). VideoA11y: Method and dataset for accessible video description. arXiv:2502.20480.

[2] Reza, M. K., Prater-Bennette, A., & Asif, M. S. (2024). Robust multimodal learning with missing modalities via parameter-efficient adaptation. arXiv:2310.03986v3.

[3] Gao, Y., Fischer, L., Lintner, A., & Ebling, S. (2024). Audio description generation in the era of LLMs and VLMs: A review of transferable generative AI technologies. arXiv:2410.08860.

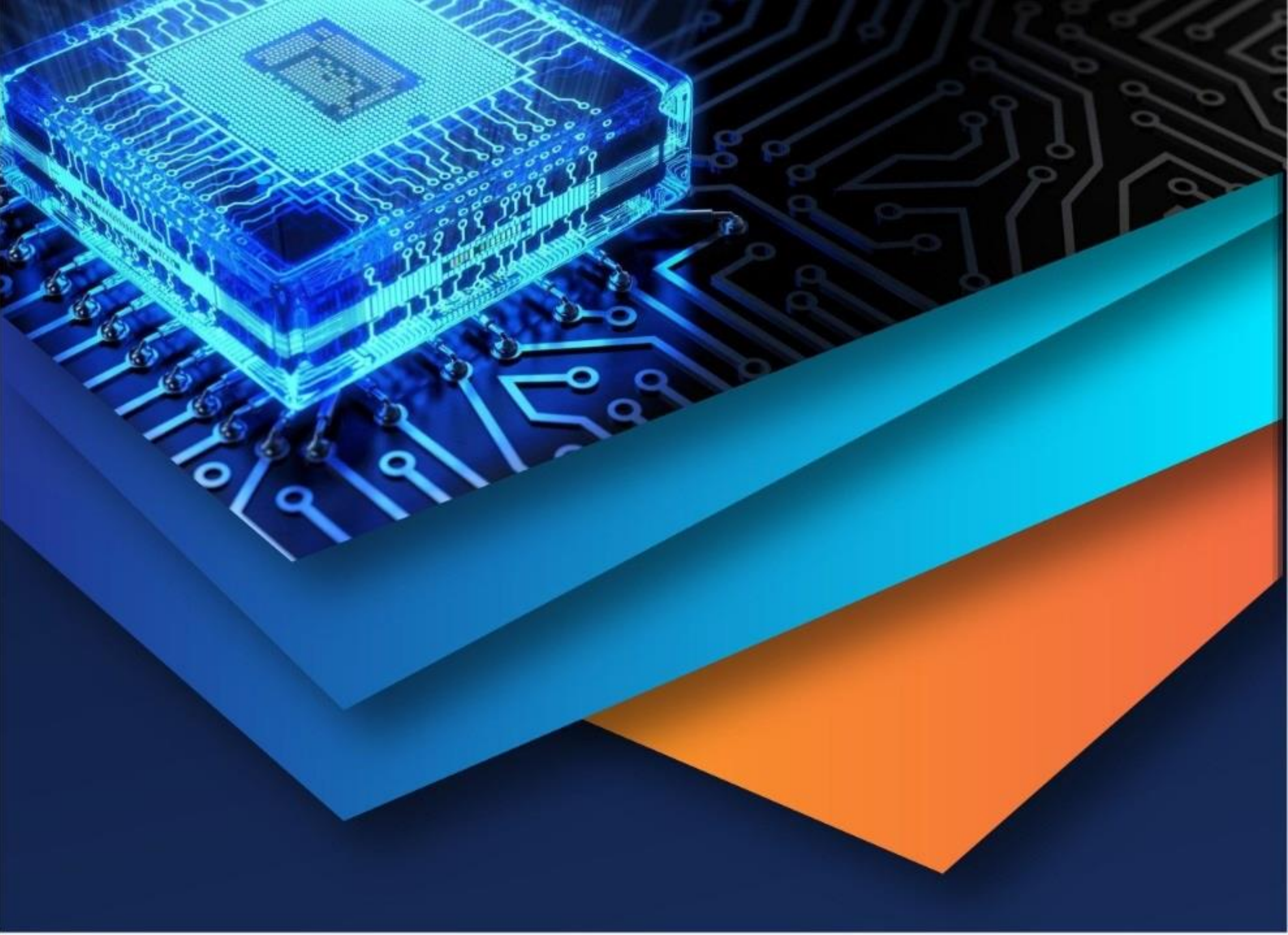
[4] Wang, X., Zheng, Y., Zhang, R., Zhang, Y., Zhou, J., Zhou, B., & Liu, Z. (2025). NarrAD: Automatic generation of audio descriptions for movies with rich narrative context. IEEE Transactions on Multimedia.

[5] tho Pesch, P., Bouqueau, R., & Montagud, M. (2020). White paper: Recommendations for immersive accessibility services. ImAc Project H2020.

[6] Soldan, M., Aradhye, H., Chen, X., Hidary, J., Holness, G., Huang, Z., ... & Malik, J. (2022). DistinctAD: Distinctive audio description generation in contexts. In CVPR 2024.

[7] Lin, J., He, C., Zeng, A., Wang, H., Zhang, Y., Yu, J., ... & Ding, X. (2023). Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In EMNLP 2023 Demo Track.

[8] Wang, J., Xu, J., Gao, Y., Hu, Q., Jiang, Y., & Chen, Y. (2024). InternVideo2: Scaling foundation models for multimodal video understanding. arXiv:2403.15377v4.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)