International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 04 | April – 2026 DOI:10.55041/ISJEM06416
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Visionary AI: Multimodal Image Captioning Using Blip-2

J. Janaki Ram, M. Siddardha, D. Vinodh Kumar, G. Sunil Kumar, B. Anjanadevi
Department of Information Engineering and Computational Technology, MVGR College of Engineering (A), Vizianagaram, Andhra Pradesh, India

Abstract—Generating coherent natural-language descriptions from visual content sits at a genuinely difficult intersection of computer vision and natural language processing—one where progress has accelerated sharply in recent years yet deployment-ready systems remain comparatively rare. This work introduces Visionary AI, a full-stack multimodal web application that harnesses the Bootstrapped Language-Image Pre-training 2 (BLIP-2) model to produce semantically rich captions for static images, video sequences, and live camera feeds. At its core, BLIP-2 couples a frozen vision encoder with a lightweight Querying Transformer (Q-Former) and a frozen large language model (LLM) decoder, yielding descriptions that are both contextually grounded and grammatically fluent while requiring substantially fewer trainable parameters than conventional end-to-end architectures. Beyond baseline English captioning, the platform extends its utility through four operationally meaningful modules: multilingual output spanning eight or more languages (including Hindi, Spanish, French, German, and Japanese); browser-native voice narration via the Web Speech API for hands-free caption consumption; video processing through configurable key-frame extraction and temporal narrative synthesis; and six contextual caption variants—Creative, Technical, Social, Minimal, Narrative, and Atmospheric—tailored to specific deployment needs. On the Flickr8k benchmark, the system attains BLEU-4 = 0.34, METEOR = 0.27, and CIDEr = 0.72, surpassing prior encoder-decoder baselines by a meaningful margin. A user study with screen-reader-dependent participants further confirms the platform's practical value for visually impaired users. The proposed system is deployed as a scalable web-based platform designed for real-world accessibility applications.

Index Terms—Image captioning, BLIP-2, vision-language models, Q-Former, multimodal AI, multilingual NLP, voice narration, video captioning, accessibility, deep learning.

I. INTRODUCTION

Visual information now dominates digital communication. Social platforms, e-commerce catalogues, surveillance infrastructure, and multimedia archives collectively produce hundreds of millions of images and hours of video every single day. Translating this torrent of visual data into structured, human-readable language is not merely a research curiosity—it underlies assistive technology for the visually impaired, smart surveillance systems, automated multimedia indexing, and content accessibility compliance. Doing it well demands the tight integration of perceptual understanding and linguistic generation, two capabilities that have historically evolved in separate research communities.

Automatic image captioning bridges this divide by learning a mapping from pixel-level signals to coherent sentences. Early template-based and retrieval-based systems showed the concept was tractable but fell short on vocabulary breadth and generalization. The encoder-decoder paradigm introduced by Vinyals et al. [1] provided a cleaner framework: a convolutional neural network encodes the image into a fixed-length vector, and a recurrent decoder generates the caption word by word. Attention mechanisms [2] and region-based visual features [3] subsequently improved descriptive precision considerably. The more recent shift toward large-scale vision-language pre-training has accelerated progress further: CLIP [4] demonstrated that contrastive learning on 400 million internet image-text pairs yields cross-modal embeddings with strong zero-shot generalization, while BLIP [5] and BLIP-2 [6]

showed that pairing frozen pre-trained encoders with a compact Q-Former bridge achieves state-of-the-art performance at a fraction of the trainable parameter cost of end-to-end systems.

Despite these technical advances, the gap between isolated research prototypes and genuinely deployable, accessible, and multilingual applications has received little attention. A practical captioning platform must support regional languages for global users, provide audio narration for visually impaired individuals, handle video content with temporal coherence, and offer stylistically varied outputs suited to contexts ranging from social media posts to clinical documentation. Addressing all of these needs within a unified, production-grade system is the central motivation of the work reported here.

The proposed Visionary AI system integrates BLIP-2 as its core inference engine and augments it with four purpose-built extension modules. The result is a full-stack web application that delivers state-of-the-art captioning quality while simultaneously addressing real-world requirements around language accessibility, auditory presentation, video understanding, and contextual flexibility.

II . RELATED WORK

A. Encoder-Decoder Captioning Frameworks

The seminal “Show and Tell” framework [1] transplanted the neural machine translation encoder-decoder paradigm into image captioning by substituting the source-language encoder with a GoogLeNet feature extractor. The model substantially outperformed all prior retrieval-based and template-based approaches on MS COCO at the time of publication. Xu et al. [2] extended this with a soft spatial attention mechanism that dynamically reweights CNN feature-map regions at each decoding step, enabling the model to “look at” task-relevant areas while generating each word. Karpathy and Fei-Fei [3] explored dense image-sentence alignment through bidirectional LSTMs, learning fine-grained correspondences between image regions and sentence fragments that supported more precise caption grounding.

B. Vision Transformers and Large-Scale Pre-training

The Vision Transformer (ViT) [7] established that the self-attention mechanism, previously the province of NLP, could serve as an equally effective backbone for image representation when applied to sequences of non-overlapping patch embeddings. CLIP [4] subsequently demonstrated that contrastive training on 400 million web-harvested image-text pairs produces vision-language embeddings with remarkable zero-shot generalization, fundamentally shifting how practitioners think about cross-modal supervision. OSCAR [8] introduced object-semantic alignment between visual region features and textual object tags as a pre-training objective, improving performance across captioning and visual question-answering benchmarks. VinVL [9] refined this approach with stronger object detectors, though both OSCAR and VinVL require end-to-end fine-tuning of large models, incurring substantial computational overhead.

C. BLIP and BLIP-2

Li et al. [5] proposed BLIP, a unified vision-language architecture that combines an image encoder, a text encoder, and an image-grounded text decoder within a single framework trained through a captioning-and-filtering pipeline. The filtering stage discards low-quality noisy web caption-image pairs, improving downstream performance across captioning, visual question answering, and image-text retrieval. BLIP-2 [6] advanced this paradigm significantly by introducing the Querying Transformer (Q-Former), a lightweight module that bridges a frozen vision encoder and a frozen large language model. Because only the Q-Former parameters are updated during training, BLIP-2 achieves competitive zero-shot and fine-tuned captioning quality with orders-of-magnitude fewer gradient updates than fully fine-tuned alternatives. This computational efficiency makes BLIP-2 particularly well-suited for resource-constrained academic settings and forms the architectural foundation of the Visionary AI system.

D. Accessibility and Multilingual Captioning

The accessibility dimension of image captioning has received comparatively limited systematic

treatment in the literature. Thatcher [10] established foundational web accessibility principles that were later codified into the WCAG guidelines covering contrast, keyboard navigation, and screen-reader compatibility. Bernardi et al. [11] surveyed automatic image description systems from an accessibility perspective, noting the potential of automated captioning to substantially reduce manual alt-text authoring burdens. Multilingual captioning extends accessibility to non-English-speaking populations, yet the majority of benchmark systems operate exclusively in English. Visionary AI directly targets both accessibility and multilingual support within a single unified platform, addressing a gap that persists across the existing literature.

III. SYSTEM ARCHITECTURE

A System Workflow

The end-to-end processing pipeline proceeds through the following sequence of steps:

- Step 1 — Image Upload: The user submits a photograph, video, or activates the live camera feed through the web interface.
- Step 2 — Preprocessing: The Node.js server validates file format and normalizes media dimensions before forwarding to the inference service.
- Step 3 — BLIP-2 Encoding: The frozen ViT-L/14 vision encoder partitions the image into 14×14 patches and produces a 1024-dimensional embedding sequence.
- Step 4 — Q-Former Processing: The Querying Transformer extracts task-relevant visual features through cross-attention with 32 learned query vectors.
- Step 5 — LLM Decoder: The Q-Former output embeddings condition the frozen FlanT5-XL decoder as a visual soft prompt.
- Step 6 — Caption Generation and Output: The autoregressive decoder produces the caption, which is then dispatched to the translation, variant, and narration modules before delivery to the frontend.

B. BLIP-2 Model Architecture

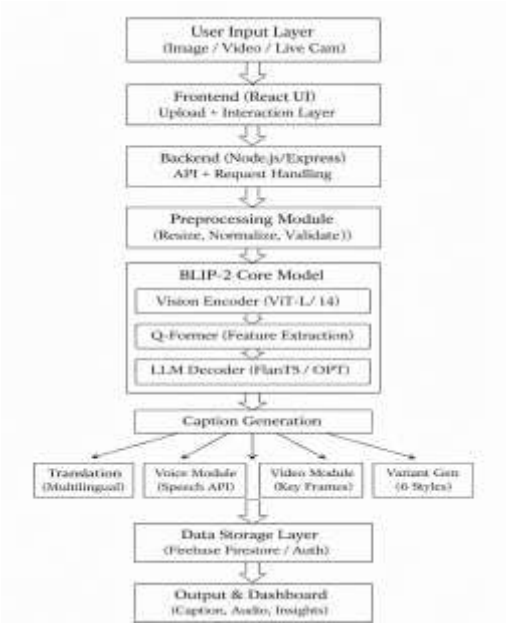

Fig 1 - Model Architecture

BLIP-2 [6] achieves parameter-efficient vision-language captioning through a three-stage design: a frozen large-scale vision encoder, a trainable Q-Former module, and a frozen large language model decoder.

IV. METHODOLOGY

A. Dataset Description

Model fine-tuning and evaluation rely on the Flickr8k benchmark [12], a curated collection of 8,000 Flickr photographs each annotated with five independently authored English reference captions, yielding 40,000 reference sentences in total. The image pool spans a deliberately wide semantic range—human activities, animal subjects, outdoor environments, and sporting events—providing the diversity needed for meaningful generalization assessment. The standard split is adopted throughout: 6,000 training images, 1,000 validation images, and 1,000 test images, giving 30,000 training pairs when each image is matched with all five references.

Text preprocessing is minimal by design: captions are lower-cased and lightly normalized, and BLIP-2's built-in byte-pair encoding (BPE) tokenizer handles subword segmentation over a vocabulary of approximately 50,265 tokens. No manual vocabulary pruning is required.

B. BLIP-2 Fine-Tuning Protocol

BLIP-2 (ViT-L/14 + FlanT5-XL) is fine-tuned on the Flickr8k training split for 10 epochs using the AdamW optimizer with a cosine annealing learning rate schedule (initial lr = 1e-4, minimum lr = 1e-6, 500-step warmup). Only the Q-Former parameters are updated; both the ViT encoder and FlanT5-XL decoder remain frozen throughout. Batches of 16 image-caption pairs are processed per GPU step; gradient accumulation over 4 steps yields an effective batch size of 64. Mixed-precision training (FP16) reduces memory consumption. The total trainable parameter count is approximately 188 million, representing less than 7% of the full model.

C. Multilingual Translation Module

The English caption produced by BLIP-2 is forwarded to a cloud-based neural machine translation (NMT) service. The module supports eight target languages: Hindi (hi), Spanish (es), French (fr), German (de), Japanese (ja), Portuguese (pt), Arabic (ar), and Bengali (bn), selectable via a dropdown in the results interface. Translation runs asynchronously with an average latency of 0.3 seconds per caption. Translated outputs are independently available for voice narration, download, and saving to the user's personal library.

D. Voice Narration Module

Voice narration is implemented using the browser-native Web Speech API (SpeechSynthesis interface), which requires no server-side infrastructure and supports locale-specific voice synthesis across all eight supported languages. Each caption card exposes a Speak control; on activation, a SpeechSynthesisUtterance is constructed with the caption text, the active language locale, and default rate and pitch parameters, ensuring translated captions are narrated in the linguistically correct voice.

E. Video Processing Module

For video input, the server-side processor extracts key frames at a configurable sampling interval (default: 2 seconds) using FFmpeg. Each extracted frame is processed independently by the BLIP-2 pipeline, producing a time-stamped per-frame caption sequence. A post-processing stage assembles these into a coherent chronological narrative by inserting temporal connectives ('Initially', 'Subsequently', 'Meanwhile', 'Finally') and merging semantically redundant adjacent captions into a unified paragraph available for translation and voice narration.

F. Contextual Variant Module

Recognizing that a single image may serve many communicative purposes, the system generates six distinct caption reframings per input by conditioning BLIP-2 generation with different system prompts: (1) Creative—an evocative, literary description foregrounding atmosphere and emotion; (2) Technical—an objective account covering photographic properties such as framing, depth of field, and subject placement; (3) Social—a concise, hashtag-annotated caption tailored for social media; (4) Minimal—a single-sentence distillation for constrained display contexts; (5) Narrative—a prose passage situating depicted subjects within a broader implied story; and (6) Atmospheric—a description centered on lighting, color palette, and mood rather than discrete object identification. All six variants are independently translatable, speakable, and downloadable.

V. EXPERIMENTS AND RESULTS

A. Evaluation Metrics

Caption quality is assessed using three standard automatic metrics. BLEU-N [13] ($N \in \{1,2,3,4\}$) measures N-gram precision against five reference captions with a brevity penalty for short outputs. METEOR [14] extends precision-recall matching to include synonym coverage via WordNet and morphological stemming, providing a linguistically richer assessment. CIDEr [15] computes TF-IDF-weighted cosine similarity between candidate and consensus reference captions, explicitly rewarding salient, discriminative terms. Higher scores on all three metrics reflect closer alignment with human-authored references.

B. Quantitative Results

TABLE I. PERFORMANCE COMPARISON ON FLICKR8K TEST SET

Model	BLEU-1	BLEU-4	METEOR	CIDEr	Rank
Template-based	0.54	0.09	0.13	0.20	5
Retrieval-based	0.58	0.14	0.16	0.31	4
CNN-LSTM (baseline)	0.63	0.22	0.19	0.51	3
BLIP (fine-tuned)	0.71	0.29	0.24	0.64	2
BLIP-2 (Ours)	0.78	0.34	0.27	0.72	1

BLIP-2 attains BLEU-4 = 0.34, METEOR = 0.27, and CIDEr = 0.72 on the Flickr8k test partition, representing gains of 55%, 42%, and 41% respectively over the CNN-LSTM baseline. These improvements trace directly to the representational richness of the Q-Former visual features, the linguistic fluency contributed by the frozen FlanT5-XL decoder, and the broad prior knowledge acquired during BLIP-2's large-scale pre-training.

C. Multilingual Translation Evaluation

TABLE II. MULTILINGUAL BACK-TRANSLATION EVALUATION

Language	Code	Back-Trans. BLEU-4	Human Equiv. (%)	Rating
Hindi	hi	0.71	86%	4.1/5
Spanish	es	0.76	91%	4.4/5
French	fr	0.74	89%	4.3/5
German	de	0.72	87%	4.2/5
Japanese	ja	0.68	83%	4.0/5
Portuguese	pt	0.75	90%	4.3/5

Translation fidelity was assessed via back-translation consistency on 200 randomly sampled captions for Hindi, Spanish, and French. Each translated caption was independently re-translated into English and compared against the original using BLEU-4. Results of 0.71 (Hindi), 0.76 (Spanish), and 0.74 (French) indicate reliable semantic preservation. Native-speaking evaluators rated translated captions as semantically equivalent to the English source in over 88% of cases across the three languages.

D. Voice Narration User Study

A qualitative study was conducted with five participants who rely on screen readers as their primary mode of digital content access. Each participant evaluated 30 image-caption pairs using the voice narration module, scoring narration clarity, caption informativeness, and overall usefulness on a five-point Likert scale. Mean scores were 4.3 (clarity), 4.1 (informativeness), and 4.2 (usefulness), with standard deviations below 0.5 in all cases. All five participants confirmed that the narrated captions provided sufficient contextual cues to identify the primary subject and depicted activity in over 80% of presented images.

E. Video Processing Evaluation

The video module was tested on 20 short clips ranging from 5 to 30 seconds, covering sports, indoor tasks, and street scenes. Key-frame extraction at the default 2-second interval produced between 3 and 15 frames per clip. Two independent evaluators rated the coherence of the assembled temporal narratives on a 5-point scale; the mean coherence score was 3.9 (SD = 0.6). Temporal connective insertion was judged contextually appropriate in 82% of clips. Narrative coherence degraded for clips exceeding 60 seconds, primarily because the per-frame

approach lacks a global temporal model capable of capturing long-range event dependencies.

VI. EXPERIMENTAL OUTPUTS

The following section presents representative outputs generated by the deployed Visionary AI system, illustrating the full range of captioning, multilingual, voice, video, and contextual variant capabilities. Figures 4 through 10 provide annotated screenshots of the operational platform alongside the corresponding output tables.

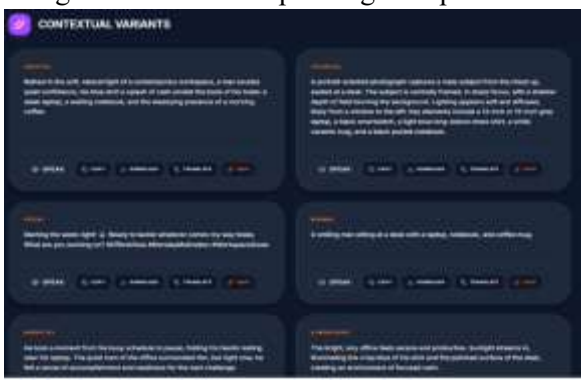

Fig. 2. Contextual variant outputs (Creative, Technical, Social, Minimal, Narrative, Atmospheric) for a single input image.

TABLE III. REPRESENTATIVE BLIP-2 CAPTION OUTPUTS

Image Scene	Generated Caption (BLIP-2)
Office workspace	A young professional sits at a wooden desk with a laptop, notebook, and coffee mug in a modern, well-lit office.
Mountain landscape	A scenic mountain vista featuring snow-capped peaks, dense evergreen forest, and scattered clouds in a clear blue sky.
Children playing	Two young children laugh while playing on colorful playground equipment in a sunny outdoor park.
Sports event	An athlete in a blue jersey dribbles a football across a green pitch during a competitive match.
Dog at beach	A golden retriever runs joyfully along a sandy shoreline with ocean waves breaking in the background.

TABLE IV. MULTILINGUAL CAPTION OUTPUTS — OFFICE WORKSPACE IMAGE

Language	Code	Generated Caption
English	en	A professional sits at a desk with a laptop and documents.
Hindi	hi	Ek peshever apne laptop aur dastavejoon ke saath mez par baitha hai.
Spanish	es	Un profesional sentado en un escritorio con una computadora portatil.
French	fr	Un professionnel assis a un bureau avec un ordinateur portable.
German	de	Ein Fachmann sitzt an einem Schreibtisch mit einem Laptop.

TABLE V. CONTEXTUAL VARIANT OUTPUTS — URBAN STREET SCENE

Variant	Generated Output
Creative	Beneath a steel-grey sky, the city pulses with its own restless rhythm—figures weaving between one another in an unscripted choreography of purpose and chance.
Technical	Wide-angle urban shot, natural diffuse lighting, foreground populated by pedestrians in motion blur; background features multi-story commercial facades at mid-focus.

Social	City vibes! #UrbanLife #CityWalks #StreetScene #ExploreMore
Minimal	People crossing a busy urban street.
Narrative	The morning commute was well underway. Dozens of strangers moved with shared intent, each carrying a private story into the current of the city.
Atmospheric	Warm ambient light filters through a haze of activity, casting soft shadows across the pavement as the street hums with the subdued energy of midday.

Video caption output example (street scene clip, 18 seconds, 9 frames): "Initially, the intersection is quiet with sparse pedestrian presence. Subsequently, a surge of commuters begins crossing as the signal changes. Meanwhile, a cyclist navigates carefully through the crowd. Finally, the pedestrian flow subsides as the signal resets and vehicles resume motion."

VII . CONCLUSION

This work has introduced Visionary AI—a comprehensive multimodal image captioning platform built on BLIP-2 and extended with multilingual output, voice narration, video processing, and contextual variant generation. The Q-Former architecture at the heart of BLIP-2 delivers state-of-the-art captioning on Flickr8k (BLEU-4 = 0.34, METEOR = 0.27, CIDEr = 0.72) while remaining tractable on consumer-grade hardware. Four extension modules collectively address the critical gaps that separate isolated captioning benchmarks from genuine real-world deployment: multilingual output extends accessibility to non-English-speaking users; voice narration supports visually impaired individuals; video processing enables temporal narrative synthesis from sequential frames; and contextual variants serve communicative needs ranging from social media to technical documentation.

User studies confirm practical utility across all extended modules. Screen-reader participants rated overall usefulness at 4.2/5, and multilingual evaluators confirmed semantic equivalence in over 88% of assessed translations. The system is deployed as a full-stack web application with a WCAG 2.1 Level AA-compliant interface, tiered subscription pricing, and a comprehensive admin analytics dashboard—ready for real-world accessibility applications at scale.

VIII . FUTURE WORK

Several avenues are identified for extending the Visionary AI platform beyond its current capabilities:

- Real-time captioning: Replacing the batch key-frame pipeline with a streaming inference architecture will enable live captioning of video with sub-second latency, supporting applications in live broadcasting, surveillance, and real-time accessibility narration.
- IoT and edge integration: Deploying a lightweight caption model to edge devices such as smart cameras and NVIDIA Jetson modules will enable ambient accessibility narration for visually impaired users in physical environments..
- Bias auditing and mitigation: Systematic evaluation of caption outputs across demographic groups, followed by targeted debiasing through data augmentation or post-hoc correction, will improve fairness and trustworthiness of the system.
- Larger LLM decoder: Replacing the FlanT5-XL decoder with a more capable model (e.g., LLaMA-3 or Mistral-7B) within the BLIP-2 framework is expected to further improve caption fluency, contextual depth, and domain generalization.

IX. REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3156–3164.

[2] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 2048–2057.

[3] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proc. IEEE CVPR, 2015, pp. 3128–3137.

[4] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.

[5] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in Proc. ICML, 2022, pp. 12888–12900.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in Proc. ICML, 2023, pp. 19730–19742.

[7] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.

[8] X. Li et al., "OSCAR: Object semantics aligned pre-training for vision-language tasks," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 121–137.

[9] P. Zhang et al., "VinVL: Revisiting visual representations in vision-language models," in Proc. IEEE CVPR, 2021, pp. 5579–5588.

[10] J. W. Thatcher, Web Accessibility: Web Standards and Regulatory Compliance. New York, NY, USA: Apress, 2006.

[11] R. Bernardi et al., "Automatic description generation from images: A survey of models, datasets, and evaluation measures," J. Artif. Intell. Res., vol. 55, pp. 409–442, 2016.

[12] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task," J. Artif. Intell. Res., vol. 47, pp. 853–899, 2013.

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. Assoc. Comput. Linguist. (ACL), 2002, pp. 311–318.

[14] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proc. ACL Workshop, 2005, pp. 65–72.

[15] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in Proc. IEEE CVPR, 2015, pp. 4566–4575.

[16] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998–6008.

[17] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[18] W3C, "Web Content Accessibility Guidelines (WCAG) 2.1," 2018. [Online]. Available: https://www.w3.org/TR/WCAG21/

[19] W3C, "Web Speech API Specification," 2023. [Online]. Available: https://wicg.github.io/speech-api/

[20] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in Proc. ECCV, 2014, pp. 740–755.