# Generating and Adapting Audio Description with Vision–Language Models for Blind and Low-Vision Users

Audio description (AD)—spoken narration of visual content—plays a crucial role in making digital media accessible to blind and low-vision (BLV) audiences. Manual AD is costly and time-consuming, requiring trained describers, script preparation, voice recording, editing, and quality assurance. As a result, coverage has not kept pace with the massive growth of digital media, leaving much of today's online videos content inaccessible. Automated AD with vision–language models (VLMs) offers scalability, but outputs tend to be verbose, redundant, or misaligned with audio tracks. Moreover, accessibility is not one-size-fits-all: BLV users preferences vary with viewing goals and genres of the content. Meeting these challenges requires both improved baseline generation and mechanisms for personalization.

We develop a system with two components leveraging state-of-the-art VLMs (Qwen2.5-VL, Gemini 1.5 Pro, GPT-4o): a baseline generation module (**GenAD**) and an on-demand adaptation module (**AdaptAD**).

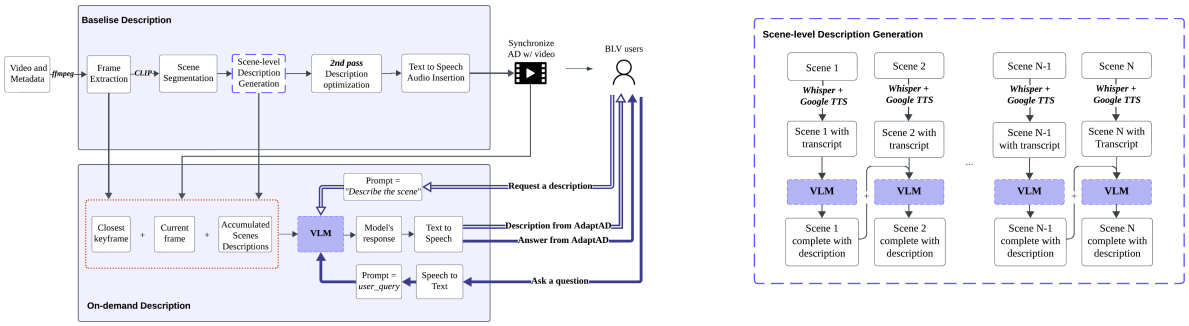


*Figure 1: End-to-end workflow of GenAD and AdaptAD*

**GenAD** pipeline begins with retrieving video and metadata (yt-dlp), extracting frames (ffmpeg), and segmenting scenes using OpenCLIP embeddings with cosine similarity. For each scene, a VLM is prompted with persona conditioning, accessibility guidelines, and contextual grounding from metadata, transcripts, and prior descriptions. Transcript alignment is improved with an ensemble of Whisper and Google Speech-to-Text, while a two-stage optimization condenses verbose drafts for inline narration and expands when extended tracks are necessary. These methods yield more concise, context-rich baselines compared to naïve prompting (Figure 2a). **AdaptAD** enables users to pause playback and request clarifications or targeted questions. Prompts build on GenAD outputs while incorporating scene transcripts and accumulated narration, producing concise, contextual responses with low latency. This reuse of GenAD content improves relevance and responsiveness, aligning with BLV users' preference for concise narration (Figure 2b).

✗ **Naive output (without guidelines → overly detailed):** A young woman with brown hair tied back stands in a vast desert landscape. She is wearing a beige, sleveless output wiht a brown belt and arm wraps. The woman looks determined as she gazes into the distance. Her expression is focused. The woman reaches for a cylindrical object attached to her belt.
➔ **With guidelines (concise but generic)**: A woman stands in a vast desert, dressed in a beige outfit. She is scanning the horizon, cylindrical object attached to her belt.
✓ **With guidelines + context (concise and specific):** Rey stands in a vast desert, scanning the horizon, lightsaber hilt at her waist.

*Figure 2a: GenAD performance*

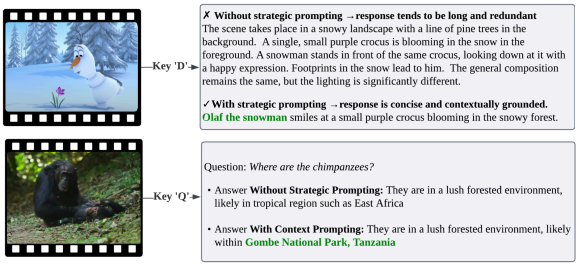


*Figure 2b: AdaptAD performance*

We evaluated the performance of GenAD on ten videos spanning entertainment, educational, and instructional genres. Seven accessibility experts rated outputs on a seven-dimension rubric (accuracy, prioritization, appropriateness, consistency, equality, delivery method, and timing/placement). Quantitative results show mean scores out of 5, with GPT-4o scoring 4.05, Gemini scoring 4.01, and Qwen scoring 3.78. Qualitative feedback described the outputs as *"good"* and *"lovely,"* noting improvements in coherence and factual accuracy under prompting, while weaknesses remained in prioritization and temporal placement. These findings show that while prompting improves baseline quality, interactive adaptation remains essential for achieving robust, user-aligned accessibility with vision–language models.