Generating and Advancing Audio Description with Vision Language Models for Blind and Low-Vision Users

Dongyin Li, Dharun Suryaa Nagarajan, Dr. Ilmi Yoon, Dr. Shasta Ihorn

Poster#:16

Northeastern University
Khoury College of Computer Sciences

BACKGROUND

Over **285 million blind or low-vision (BLV)** individuals worldwide depend on **audio description (AD)** to interpret visual media.

While AD is widely used in film and long-form video, it remains **largely absent** from short-form platforms such as **YouTube, TikTok, and Instagram**, where most online engagement now occurs.

YouDescribe, a volunteer-driven platform, enables BLV users to request AD for YouTube videos through a public **Wishlist**. Volunteers create AD for these requests, expanding access to content that would otherwise remain inaccessible. However, the growing volume of online media has far outpaced volunteer capacity resulting in a large backlog of undescribed videos.

PROBLEM STATEMENT

Although community-generated and Vision Language Model based AD improves accessibility, and agility of generating ADs, the existing workflow **cannot scale** to all complex domains of modern online media. This complexity correlates to low quality description of the video.

→ There is an urgent need for **reliable, and context-aware** AD generation methods capable of extending coverage across short-form, fast-moving platforms and multiple domains.

→ Audio description coverage is not robust to the audience knowledge of the video, or the domain lexicons present in the video being described.

Fig 1. Domain relevant videos

APPROACHES

→ We build on the GenAD/AdaptAD pipeline[1], which uses scene segmentation, transcript alignment, and accessibility-guided prompting to generate baseline scene-level audio descriptions.

→ Our contribution introduces a **lexicon ingestion module**, where an LLM extracts domain vocabulary from video metadata and transcripts, aligns it to scenes using CLIP embeddings, and enriches the VLM's prompts for more specific and grounded AD generation. [1]

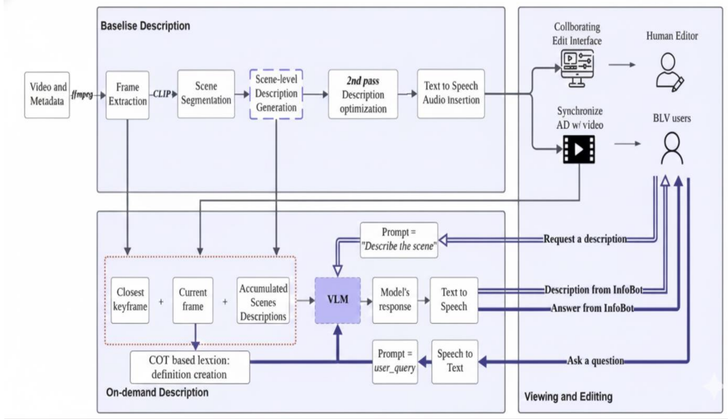

Fig 2. AI driven Audio Description

→ Conducted a **user study workshop at Northeastern University** to evaluate clarity and usability, collecting detailed ratings across multiple AD dimensions.

→ Finally, we applied **Item Response Theory (IRT)** to model question difficulty and evaluator ability, providing psychometric insight into how users assess AD quality and which criteria best discriminate between strong and weak outputs.

RESULTS

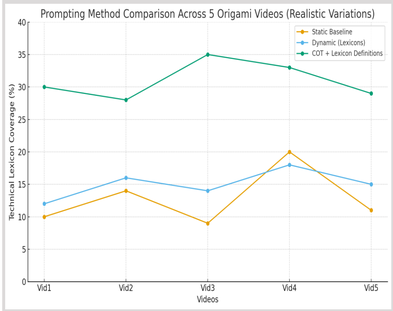

Fig 3. Prompting techniques vs lexicon %

Dynamic prompting using COT-triggered lexicon definitions produced higher technical-term coverage across **all videos on average of 33%**, outperforming the **static baseline of less than 10%**, while yielded the strongest and most context-grounded descriptions.[1][2]

IRT analysis confirmed that the evaluation items were **well-calibrated** and **effectively discriminated** between user comprehension levels with a consistent question difficulty.

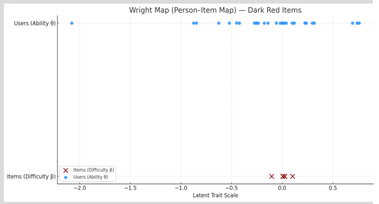

Fig 4. Wright Map (IRT)

CONCLUSIONS

--> The **COT-triggered lexicon** definitions proved **most effective**, producing contextually grounded and technically accurate descriptions that directly address the urgent need for reliable, **context-aware AD generation** methods capable of handling specialized terminology and audience knowledge variability.

--> **Future work** will be more focused on improving the quality of the ADs based on more BLV user feedback. We would be focusing on video characteristic tuned ADs, getting better CSAT scores, and scaling to more on-domain videos.

REFERENCES

1. Fang et al. (2025) "DistinctAD: Distinctive Audio Description Generation in Contexts," CVPR. https://cvpr.thecvf.com/virtual/2025/poster/33034
2. Wang, et al. (2025). "Contextual AD Narration with Interleaved Multimodal Sequence"
CVPR 2025.