Poster#: 32

Scalable and Context-Aware Audio Description via VLM for Blind and Low Vision Audiences

Lana Do[1], Dr. Ilmi Yoon[2]
[1]Align M.S. Computer Science, [2]Teaching Professor, Director of Computing Programs - Silicon Valley
Northeastern University Silicon Valley

Github Repo

Introduction

Over 285 million blind or low vision (BLV) individuals rely on audio description (AD) to access visual content. While AD is mandated for professional media, it's often missing from short-form and user-generated platforms like YouTube, Instagram, and TikTok.

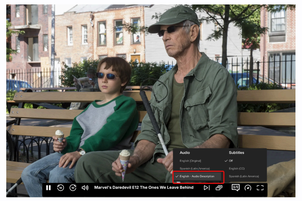

Fig.1: Daredevil is one of thousands of titles with AD on Netflix or Disney+

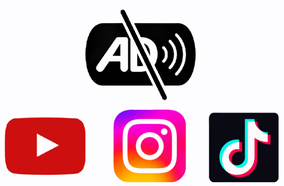

Fig.2: Social media platforms offer little to no AD support

YouDescribe is a platform that enables BLV users to request audio descriptions, which are then created by volunteers. Over 3,000 contributors have produced 6,400+ described videos using its guided, accessible interface.

Motivation

Volunteer efforts can't keep up with the growing volume of digital content—**92.5%** of YouDescribe Wishlist videos remain without descriptions. A human-in-the-loop (HITL) system offers a scalable, effective solution.

Fig. 3: Human-in-the-loop design leverages AI speed and human insight to deliver the best possible AD experience

When human-written AD isn't available, BLV users can request AI-generated drafts for volunteers to refine—making the process faster, less overwhelming, and more accessible.

Fig. 4: BLV users and describers can request AI description on any video without current AD

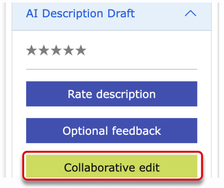

Fig. 5: Describers refine AI drafts for more polished AD with less time and effort

Automated ADs often lack context and are verbose. This project uses VLMs with context-informed prompts to generate high-quality drafts—enhancing users' experience, reducing volunteer effort, and enabling on-demand descriptions that empower BLV audiences as active participants.

Approach

This system provides automated support for generating both baseline and on-demand video descriptions. Each mode offers distinct capabilities, while also sharing a unified processing framework. The complete architecture is outlined in Fig. 6 below.

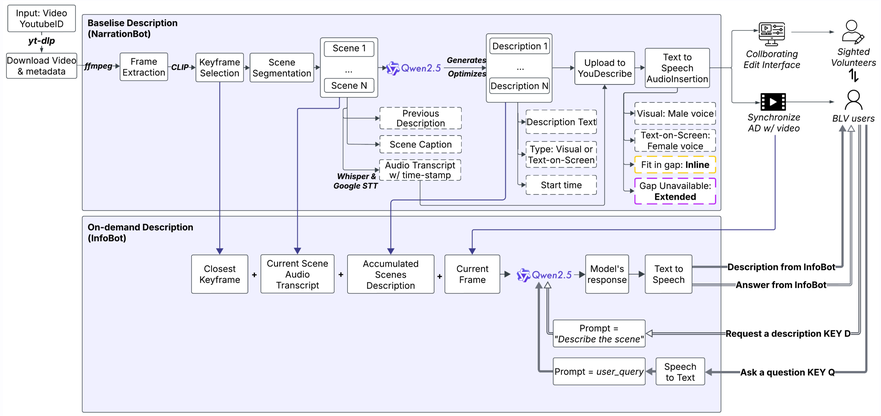

Fig. 6: End-to-end workflow for NarrationBot and InfoBot description generation.

NarrationBot generates audio descriptions of key visuals using video, transcripts, and captions. It segments scenes with **OpenCLIP**, transcribes audio with **Whisper + Google STT**, and uses **Qwen2.5-VL** for both description generation and optimization—refining drafts for clarity, timing, and quality. The final output—including the original video's audio timestamps and the optimized descriptions with their corresponding timings—is structured into a JSON file to upload to YouDescribe database and server.

InfoBot enables on-demand interaction. **Pressing 'D'** pauses the video and plays a spoken description. **Pressing 'Q'** lets users ask questions, with Qwen2.5-VL selecting the most informative frame and providing real-time, context-aware answers.

Qualitative Analysis

I. NarrationBot

A qualitative example shows how context-informed prompting improves narration flow. For instance, in a Spider-Man scene, prior context helps identify characters and events.

W/o context prompt: A young man, visibly distresseed, kneels beside an older main who lies on the ground, appears injured on the street.
With context promp: Peter kneels beside Uncle Ben, offering comfort as he lies injured on the street following the carjacking.

Fig. 7: Titled Uncle Ben's Death Vengeance Scene – Spider-Man (2002). Prior context and transcripts enhance narration accuracy, capturing names and key events.

Qualitative Analysis (cont.)

A second example: the optimization module trims a long narration to fit a natural gap—producing well-timed output that enhances the users' experience.

Raw ouput: Sweeping view of a lush, green forested landscape. Dense layers of foliage stretch across the rolling hills under a partly cloudy sky.(Extended)

Optimized: A lush forest with rolling hills appears under a partly cloudy sky. (Inline)

Raw output: A chimpanzee swings through the dense tree canopy. The surrounding foliage rustles and branches sway, as shafts of light break through the leaves.(Extended)

Optimized : Chimpanzee swings through dense tree canopy, branches sway with its movement.(Inline)

Fig. 8: Titled jane goodall. Optimization improves timing and clarity without losing meaning.

II. InfoBot

Qualitative analysis shows InfoBot answers questions accurately and provides meaningful responses by combining scene context—going beyond the current frame.

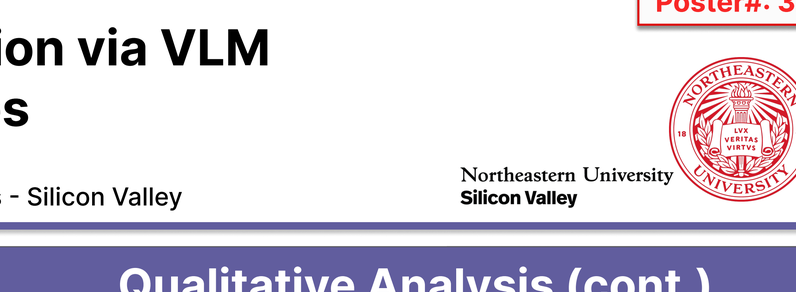

Fig. 9: InfoBot infers Uncle Ben was shot by a car thief

Fig. 10: InfoBot identifies the location as Gombe National Park

Conclusion & References

I. CONCLUSION

VLM can generate high-quality AD, but techniques like context-informed prompting and adherence to AD guidelines are key to producing content that better serves BLV users. More importantly, Human-in-the-loop refinement remains critical—better initial drafts make the process more manageable, especially for novice describers. InfoBot further enhances access by giving BLV users greater agency, allowing them to actively engage with video content rather than remain passive viewers.

II. REFERENCES

[1] Aditya Bodi, Pooyan Fazli, Shasta Ihorn, Yue Ting Siu, Andrew T. Scott, Lothar Narins, Yash Kant, Abhishek Das, and Ilmi Yoon. 2021. Automated Video Description for Blind and Low Vision Users. Conference on Human Factors in Computing Systems - Proceedings (5 2021)
[2] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., & Lin, J. (2025). Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923.