CustomQA: Customized Video Descriptions and Video Question Answering for Blind and Low Vision

Nghi Huynh, Computer Science
Mentor: Hasti Seifi, Assistant Professor
Ira A. Fulton Schools of Engineering

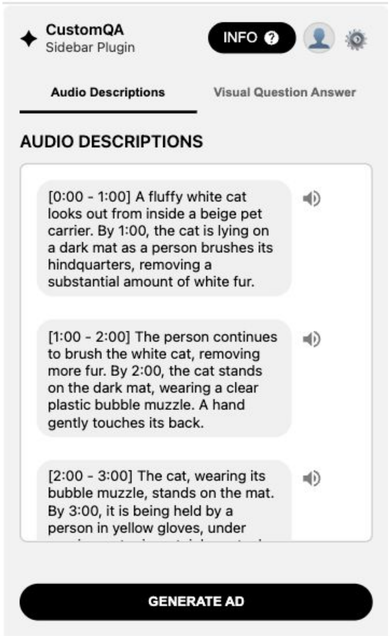
INTRODUCTION

- BLV users often consume videos through audio descriptions (AD), spoken audio describing visuals
- ADs are usually written and narrated by humans, but are not always available on online videos due to being labor-intensive.
- Previous work in AD systems are limited when accounting for unique individual needs of user [1].

Research Goals:

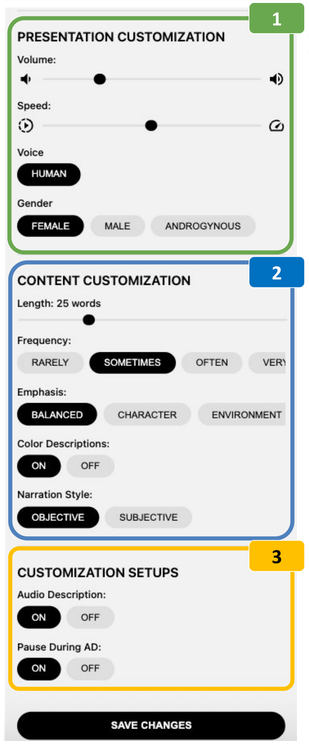
1. Develop a Chrome plugin to generate ADs and answer questions with cutomizable settings.
2. Investigate common question types BLV users ask with and without AD customizations.

AUDIO DESCRIPTION



1. **Video Input:** Video link automatically identified by plugin.
2. **Audio Descriptions Box**: Clicking ”Generate AD” prompts Gemini to create ADs based on a base prompt. If customizations are toggled, respective prompts are added to the base prompt. Generated ADs are displayed in AD Box.

AD CUSTOMIZATION SETTINGS

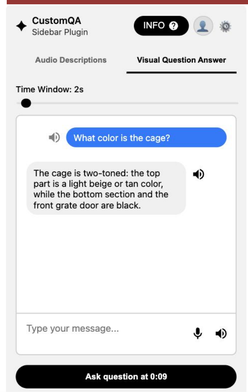


1. **Presentation Customization**: Users can adjust OpenAI Whisper text-to-speech **(1)** volume **(2)** speed **(3)** voice **(4)** gender
2. **Content Customization**: Users can adjust audio description **(1)** word length **(2)** frequency **(3)** emphasis **(4)** color **(5)** narration style.

Length keeps ADs to chosen amount of words. AD frequency allows AD generation every 1 minute (rarely), 30s (sometimes), 15s (often), or 8s (very often). Emphasis allows users to prompt Gemini to focus on character(s), environment(s), instruction(s), or balanced audio descriptions.

3. **Customization Setups**: Users can toggle on/off automatic AD playback and video pause during AD.

VISUAL QUESTION ANSWERING



1. **VQA Box**: Questions are asked via keyboard or microphone on a video’s current timestamp. Speaker button reads aloud questions and answers based on presentation settings.
2. **Content Customization**: Users can adjust **(1)** word length of question answers. If a question is not answerable, the plugin expands 3s, 9s, 30s, and 60s forward and backwards from current timestamp to add context.

USER STUDY

Participants

- 5 sighted participants, 1 female

Study Design

- Introductory training session
- 6 videos, ~2-3 min each:
  - 2 how-to, 2 vlogs, 2 entertainment
- Users watch 1 video of each category with **default settings** and **customized settings**, where the order is counterbalanced.

RESULTS

Female was more commonly chosen, and longer responses (25+ words) favored. **Default** had lower ratings of effectiveness compared to **Customized** settings.

Common Question Types (Default vs. Customized):

- More questions asked about character, colors, and presence of objects/people
- Audio-visual inference (e.g., "Who said that?")

REFERENCES

[1] Natalie, Rosiana, et al., “Audio Description Customization.” ACM SIGACCESS Conference on Computers and Accessibility (St. John’s, NL, Canada) (ASSETS ’24) (2024).
[2] Cheema, Maryam, et al., “DescribePro: Collaborative Audio Description with Human-AI Interaction.” arXiv preprint arXiv:2508.01092 (2025).
[3] Cheema, Maryam, et al., "ViDscribe: Multimodal AI for Customizing Audio Description and Question Answering in Online Videos," 2026, arXiv:2603.14662.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University