ADx3: A Collaborative Workflow for High-Quality Accessible Audio Description

Lana Do, Shasta Ihorn, Charity M Pitcher-Cooper, Juvenal Francisco Barajas,
Gio Jung, Xuan Duy Anh Nguyen, Sanjay Mirani, Ilmi Yoon

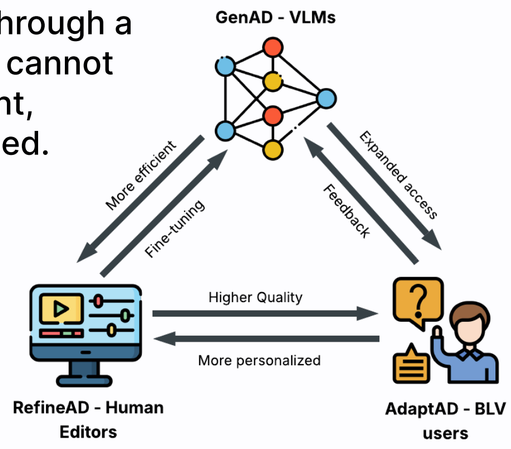
NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Over 285 million BLV individuals rely on audio description (AD), yet it is still scarce on platforms like YouTube, Instagram, and TikTok.

YouDescribe allows users to request AD through a volunteer-fulfilled Wishlist, but volunteers cannot keep pace with the growth of video content, leaving **93%** of requests remain undescribed.

ADx3, a collaborative workflow:
(1) GenAD: AI-generated drafts
(2) RefineAD: human-in-the-loop (HITL) editing interface
(3) AdaptAD: on-demand descriptions



System

Fig. 1 shows the full ADx3 system: **GenAD**, **RefineAD** and **AdaptAD** for user-driven interaction. GenAD and AdaptAD rely on VLMs, so we selected three VLMs: **Qwen2.5-VL, Gemini 1.5 Pro, and GPT-4o**.

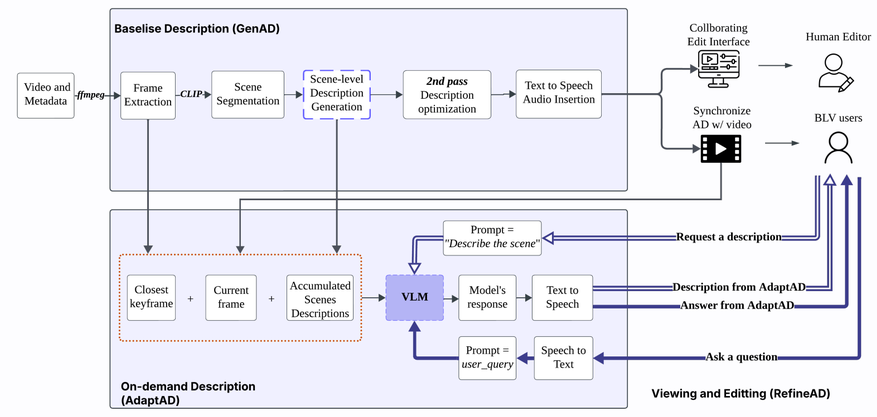


Fig.2: Complete system workflow of ADx3

GenAD: Qualitative Insights

First scenes are segmented by calculating their **OpenCLIP embeddings** exceed a certain cosine similarity thresholds. Each scene is then prompted with DCMP guidelines, metadata from the video, transcripts from **Whisper and Google STT**, and the previous scene's description for continuity. This approach improves specificity and contextual accuracy, as shown in Fig. 3.

✗ **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 with 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.

Fig. 3: Pogression from simple VLM prompting to guideline-informed and contextual prompting

GenAD: Quantitative Results

Seven accessibility experts rated three anonymized AI descriptions for ten videos on seven dimensions: ***Accurate, Prioritized, Appropriate, Consistent, Equal, Delivery Method, and Timing & Placement***, using a 1–5 scale (1 = Critical issues, 2 = Major Issues, 3 = Noticeable Issues, 4 = Minor Issues, and 5 = Just Right).

Criteria	Qwen		Gemini		GPT	
	Mean	SD	Mean	SD	Mean	SD
Overall	3.78	1.00	4.01	1.02	**4.05**	0.97
Accurate	3.63	1.13	**3.94**	1.08	**3.94**	1.13
Prioritized	3.43	1.06	3.61	1.11	**3.76**	1.10
Appropriate	3.90	1.06	**4.16**	1.08	**4.16**	1.06
Consistent	4.03	0.97	**4.31**	1.02	4.27	0.96
Equal	4.41	0.95	**4.56**	0.88	4.41	1.06
Delivery Method	3.37	1.01	3.67	1.15	**3.84**	0.98
Track Placement	3.69	0.94	3.79	1.01	**3.97**	0.87

Table 1: Overall and per-criteria AI scores with separate Mean and Standard Deviation (SD). Bold indicates the top-performing model per row

Experts rated AI-generated AD higher on dimensions such as equal, consistent, and appropriate, and described some tracks as *"good"* or *"lovely,"* but still noted prioritization and timing issues.

RefineAD

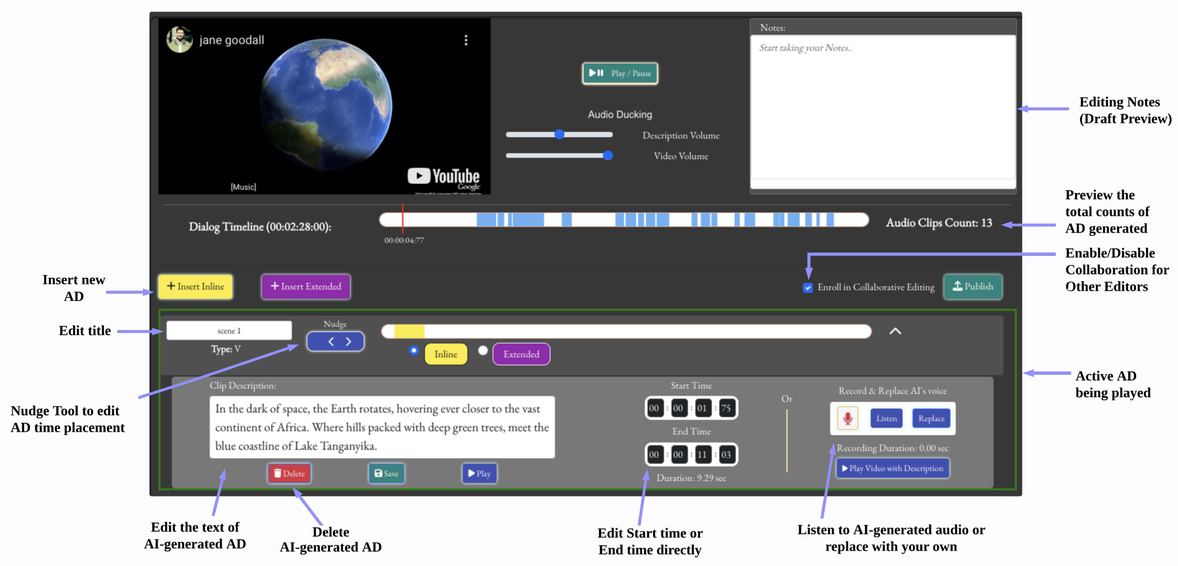


Fig. 4: RefineAD interface for editors. Screen reader compatibility allows BLV users to access and edit as well

RefineAD gives users agency to shape AI drafts into descriptions that reflect their own voices. Its accessible tools make it easy to adjust timing, wording, and delivery style.

AdaptAD

AdaptAD builds on GenAD's prompting framework, using scene transcripts and accumulated descriptions to generate concise, contextually relevant responses.

Key 'D'

✗ **Without strategic prompting → response tends to be long and redundant**
The scene takes place in a snowy landscape with a line of pine trees in the background. A single, small purple crocus is blooming in the snow in the foreground. A snowman stands in front of the same crocus, looking down at it with a happy expression. Footprints in the snow lead to him. The general composition remains the same, but the lighting is significantly different.

✓ **With strategic prompting → response is concise and contextually grounded.**
Olaf the snowman smiles at a small purple crocus blooming in the snowy forest.

Fig. 5: Strategic prompting refines verbose output into concise and specific character name.

Key 'Q'

Question: *Where are the chimpanzees?*

- Answer **Without Strategic Prompting:** They are in a lush forested environment, likely in tropical region such as East Africa
- Answer **With Context Prompting:** They are in a lush forested environment, likely within **Gombe National Park, Tanzania**

Fig. 6: Without context prompting, the model gives a vague regional response; With context prompting, it correctly specifies the detailed location.

Conclusions

- ADx3 combines AI-generated baselines (GenAD), human-in-the-loop refinement (RefineAD), and adaptive user control (AdaptAD) to support scalable, high-quality audio description.
- Accessibility specialists rated AI-generated descriptions as *"good"* but not *"excellent,"* noting persistent gaps in prioritization, choice between inline/extended description, timing and placement.
- RefineAD adds the human judgment and contextual nuance that GenAD lacks, addressing these quality gaps. AdaptAD personalizes narration for diverse BLV preferences by letting users request more or less detail and ask specific questions.
- Together, ADx3 forms a **dynamic, collaborative** process that expands access while centering both describers (novice and experienced) and BLV users' agency and participation.

Future Work

- Expand evaluation with BLV users across more videos and genres.
- Refine AD quality criteria, especially for timing and delivery.
- Integrate more expressive TTS to improve emotional clarity.
- Use human feedback to personalize narration style and timing.

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] Abigale Stangl, Shasta Ihorn, Yue Ting Siu, Aditya Bodi, Mar Castanon, Lothar D. Narins, and Ilmi Yoon. 2023. The Potential of a Visual Dialogue Agent In a Tandem Automated Audio Description System for Videos. ASSETS 2023- Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (10 2023)++