VideoA11y-40K: A Large-Scale Dataset for Accessible Video Understanding

Chaoyu Li, Sid Padmanabhuni, Maryam Cheema, Hasti Seifi, Pooyan Fazli
Arizona State University
{chaoyuli, spadma20, mcheema2, hasti.seifi, pooyan}@asu.edu

Abstract

Video descriptions are crucial for blind and low vision (BLV) users to access visual content. Current artificial intelligence models fall short because training datasets contain human annotations that do not fully meet BLV users' needs. We introduce VideoA11y, an approach that leverages multimodal large language models and video accessibility guidelines to generate descriptions tailored for BLV individuals. Using this method, we curated VideoA11y-40K dataset with 40,000 videos. Our experiments showed that VideoA11y descriptions outperform novice human annotations and are comparable to trained human annotations. Models fine-tuned on VideoA11y-40K produce high-quality accessible descriptions. Code and dataset are available at https://people-robots.github.io/VideoA11y/.

1. Introduction

The explosive growth of online video has widened the accessibility gap for BLV audiences because producing high-quality audio descriptions (AD) is expensive and slow. We introduce VideoA11y [1]—an approach that combines 42 professional AD guidelines with multimodal large language models (MLLMs)—and VideoA11y-40K, the largest dataset featuring descriptions tailored to BLV user needs. Five user studies show that VideoA11y's descriptions outperform novice human annotations and are comparable to those of trained describers in terms of descriptiveness, objectivity, accuracy, and clarity. Two complementary benchmarks confirm that both (i) plug-and-play use of VideoA11y on MLLMs and (ii) fine-tuning MLLMs on VideoA11y-40K boost accessibility, demonstrating scalable paths to high-quality video access for BLV viewers.

2. VideoA11y

VideoA11y combines 42 professional AD guidelines with MLLMs to generate BLV-focused video descriptions. As Fig. 1 shows, it extracts keyframes and, if available, includes existing human annotations, merges these with the curated AD guidelines into one prompt, then drives the MLLM to generate an accessibility-compliant description.

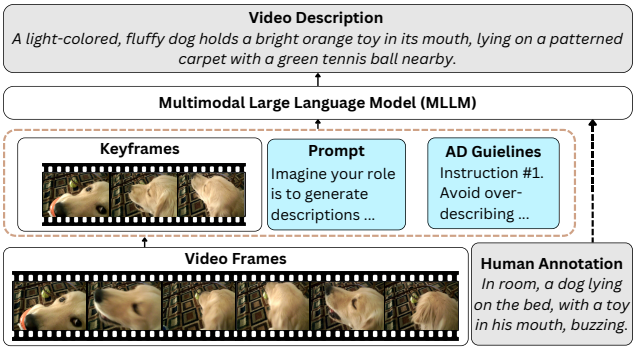

Figure 1. Overview of the VideoA11y pipeline.

Curating AD guidelines. We gathered 154 guidelines from Netflix, Ofcom, Media Access Canada, and DCMP resources. We remove duplicates and items that only focus on audio or context, then we synthesized the remaining items into 42 prompt-friendly rules.

Keyframe extraction. To capture salient visual changes, we convert frames to LUV space, compute frame-to-frame differences, smooth them with a 15-frame sliding window, and select local maxima as keyframes.

Prompt design and description generation. The extracted keyframes, optional human annotations, and the compliant prompt with 42 AD guidelines are fed to an MLLM, which generates or refines a single, accessibility-compliant description for the video.

3. VideoA11y-40K

We ran VideoA11y with GPT-4V as MLLM on three video sets—VALOR32K, VATEX, and YouCook2—to generate BLV-oriented descriptions. The result is VideoA11y-40K: 40,000 videos split into 32K/4K/4K for train/val/test set and covering 15 video content categories (Fig. 2). The average description length in the VideoA11y-40K dataset is 52.30 words, which is considerably longer than 20.30 words in the original datasets. To verify categorization accuracy, we sampled 5 videos from each category and recruited 225 MTurk participants to rate the correctness. Our results confirmed that 96% of the videos were accurately categorized.

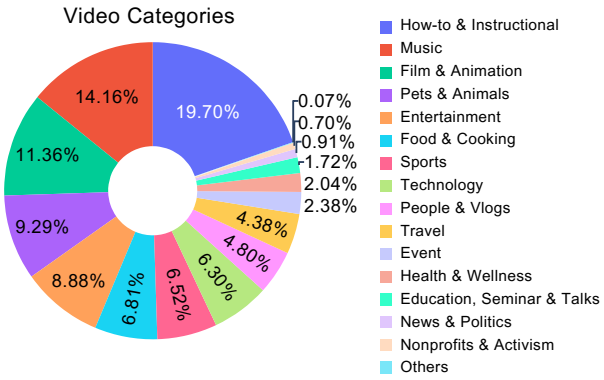

Figure 2. Distribution of video categories in VideoA11y-40K.

4. Experiments

Table 1. Zero-shot comparison of standard NLP metrics for different models with and without VideoA11y on a held-out test set. **Bold** numbers indicate better performance. VA: VideoA11y.

Model	VA	Bleu_4	ROUGE_L	CIDEr	SPICE
VILA1.5	×	4.08	20.15	8.87	17.72
VILA1.5	✓	**4.57**	**21.68**	**10.97**	**20.80**
LLaVA-Video	×	4.93	22.54	3.06	19.49
LLaVA-Video	✓	**5.34**	**23.50**	**3.21**	**22.14**

Table 2. Zero-shot comparison of custom metrics for different models with and without VideoA11y on a held-out test set. **Bold** numbers indicate better performance. VA: VideoA11y.

Model	VA	Descriptive	Objective	Accurate	Clear
VILA1.5	×	2.35	3.21	2.45	2.87
VILA1.5	✓	**2.38**	**3.48**	**2.52**	**3.02**
LLaVA-Video	×	3.68	3.34	2.76	3.67
LLaVA-Video	✓	**3.91**	**3.39**	**2.94**	**3.95**

To evaluate our approach and dataset, we used four custom metrics—descriptive, objective, accurate, and clear—to assess the overall quality of video descriptions. Additionally, we conducted both human user studies and model benchmarking. First, five user studies with 347 sighted participants, 7 professional audio describers, and 40 BLV users validated the quality of video descriptions from VideoA11y. In parallel, we benchmarked state-of-the-art (SOTA) open-source MLLMs through two settings: (1) evaluating zero-shot application of VideoA11y, and (2) evaluating models fine-tuned on the VideoA11y-40K dataset.

Human Studies. In Study 1, 150 MTurk participants rated descriptions for 150 videos across four VideoA11y variants, and the variant using GPT-4V as the MLLM with human annotations as references achieved the highest and most consistent quality scores. In Study 2 with another 150 sighted participants, we directly compared VideoA11y descriptions with those produced by novice human describers and found that VideoA11y outperformed novice annotations across all evaluation metrics. Studies 3 and 4 extended the comparison to trained human describers, with both 47 MTurk participants and seven professional audio describers preferring VideoA11y for its engaging narratives, detailed visuals, and overall clarity. Finally, in Study 5, we engaged 40 BLV users across five video categories, who consistently favored VideoA11y over novice human annotations for its greater clarity, richer details, and stronger user satisfaction.

Benchmarking VideoA11y and VideoA11y-40K. First, we applied VideoA11y to VILA1.5-40B [2] and LLaVA-Video-32B [3], and evaluated their performance on the VideoA11y-40K test set. Second, we fine-tuned LLaVA-Video-32B on VideoA11y-40K, resulting in the VideoA11y-32B model. All evaluations were conducted using standard NLP metrics and our four custom metrics. Results show that VideoA11y enables quality improvements for existing models (Tab. 1, Tab. 2), while VideoA11y-40K supports efficient training of models that generate high-quality descriptions for BLV users (Tab. 3, Tab. 4).

Table 3. Post–fine-tuning comparison of standard NLP metrics for different models on a held-out test set.

Model	Bleu_4	ROUGE_L	CIDEr	SPICE
VILA1.5	4.57	21.68	10.97	20.80
LLaVA-Video	5.34	23.50	3.21	22.14
VideoA11y-32B	**13.95**	**31.46**	**40.29**	**29.20**

Table 4. Post–fine-tuning comparison of custom metrics for different models on a held-out test set.

Model	Descriptive	Objective	Accurate	Clear
VILA1.5	2.38	3.48	2.52	3.02
LLaVA-Video	3.91	3.39	2.94	3.95
VideoA11y-32B	**3.98**	**3.94**	**3.06**	**3.97**

Acknowledgments. This research was supported by the National Eye Institute (NEI) of the National Institutes of Health (NIH) under award number R01EY034562.

References

[1] Chaoyu Li, Sid Padmanabhuni, Maryam Cheema, Hasti Seifi, and Pooyan Fazli. Videoa11y: Method and dataset for accessible video description. In *CHI*, 2025. 1

[2] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 2

[3] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 2