VideoA11y: Method and Dataset for Accessible Video Description

Chaoyu Li
School of Computing and Augmented
Intelligence
Arizona State University
Tempe, Arizona, USA
chaoyuli@asu.edu

Sid Padmanabhuni
School of Computing and Augmented
Intelligence
Arizona State University
Tempe, Arizona, USA
spadma20@asu.edu

Maryam Cheema
School of Computing and Augmented
Intelligence
Arizona State University
Tempe, Arizona, USA
mcheema2@asu.edu

Hasti Seifi
School of Computing and Augmented
Intelligence
Arizona State University
Tempe, Arizona, USA
hasti.seifi@asu.edu

Pooyan Fazli
School of Arts, Media and
Engineering
Arizona State University
Tempe, Arizona, USA
pooyan@asu.edu

Abstract

Video descriptions are crucial for blind and low vision (BLV) users to access visual content. However, current artificial intelligence models for generating descriptions often fall short due to limitations in the quality of human annotations within training datasets, resulting in descriptions that do not fully meet BLV users' needs. To address this gap, we introduce VideoA11y, an approach that leverages multimodal large language models (MLLMs) and video accessibility guidelines to generate descriptions tailored for BLV individuals. Using this method, we have curated VideoA11y-40K, the largest and most comprehensive dataset of 40,000 videos described for BLV users. Rigorous experiments across 15 video categories, involving 347 sighted participants, 40 BLV participants, and seven professional describers, showed that VideoA11y descriptions outperform novice human annotations and are comparable to trained human annotations in clarity, accuracy, objectivity, descriptiveness, and user satisfaction. We evaluated models on VideoA11y-40K using both standard and custom metrics, demonstrating that MLLMs fine-tuned on this dataset produce high-quality accessible descriptions. Code and dataset are available at https://people-robots.github.io/VideoA11y/.

CCS Concepts

• **Human-centered computing** → **Accessibility technologies;**
Accessibility systems and tools.

Keywords

Video Accessibility, Video Description, Video Understanding, Blind
and Low Vision Users, Multimodal Large Language Models

1 Introduction

New video content is created at an astounding rate, further widening the digital accessibility (a11y) gap experienced by blind and low vision (BLV) people. Video description, also known as audio description (AD), can make videos accessible to BLV users by narrating the visual content of a scene, such as actions, characters, scene changes, and interactions [6, 7, 12, 85]. For professionally created media, such as films and television shows, producing ADs requires significant collaborative efforts from a team of experts, including producers, audio description writers, voice actors, and audio engineers [24]. Thus, smaller studios and independent films may not always provide AD. For user-generated content, which has surged in popularity on platforms such as YouTube and TikTok, the implementation of ADs lags considerably behind [5]. YouDescribe [83] is an online platform where users can record and upload descriptions for YouTube videos. However, most videos in its wish list remain undescribed since the time, training, and confidence needed to create quality descriptions can deter potential contributors [53, 85]. Given the rapid increase in online videos, human description alone is insufficient, making artificial intelligence (AI)-generated audio descriptions a viable alternative.

In recent years, advances in computer vision and natural language processing (NLP) have enabled the development of new techniques for automatically generating video descriptions [18] using multimodal large language models (MLLMs) [43, 46]. These models are typically trained on general video description datasets, which include videos paired with descriptions or annotations (we use the terms 'description' and 'annotation' interchangeably) created by either humans or AI. However, existing datasets are insufficient for generating descriptions that effectively support BLV individuals in understanding video content. A key limitation is that annotations in these datasets often contain errors and fail to adhere to AD guidelines for accessibility. Human-generated descriptions also
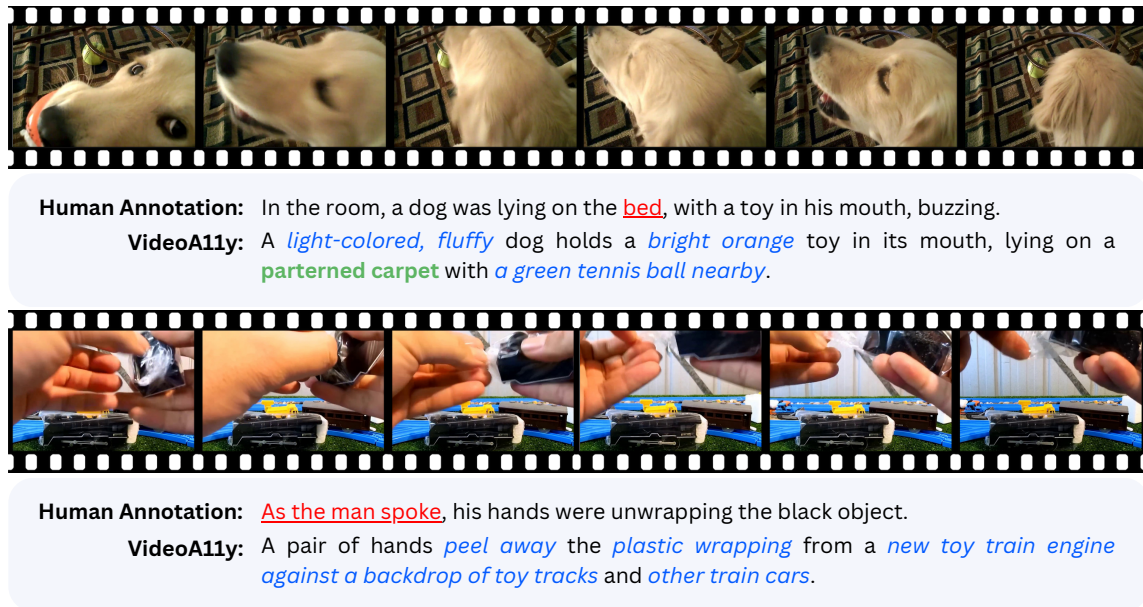
Figure 1: The human annotations and descriptions generated by VideoA11y for six consecutive frames of two sample video clips. Red underline indicates the errors in human annotations, green bold indicates the corrected facts, and *blue italics* indicates additional details.

tend to be brief and can contain grammatical, spelling, or semantic errors, which may limit video comprehension for BLV audiences.

To address this gap, we introduce the VideoA11y method and the VideoA11y-40K dataset. VideoA11y is a novel approach designed to generate high-quality descriptions from scratch or enhance existing annotations for a wide range of video categories. It aims to produce detailed and accurate descriptions, thereby improving content accessibility for BLV users. To this end, we have compiled and summarized 42 guidelines from professional AD sources that capture the needs of BLV individuals. We then leveraged MLLMs to generate accessible descriptions using a prompt that adheres to these guidelines (i.e., compliant prompt). Figure 1 shows examples of human annotations and revised descriptions generated by VideoA11y using GPT-4 Vision (GPT-4V) [62] as the MLLM. We used VideoA11y to curate VideoA11y-40K, the largest and most comprehensive video description dataset for training accessible models. The dataset includes 40,000 videos across 15 categories, all specifically described for BLV individuals. To evaluate VideoA11y and VideoA11y-40K, we asked the following research questions:

(1) **RQ1.** How do VideoA11y descriptions compare in quality to those created by novice and trained human describers?

(2) **RQ2.** How do professional describers and BLV users evaluate and prefer VideoA11y descriptions compared to human descriptions?

(3) **RQ3.** Can the VideoA11y-40K dataset enhance state-of-the-art (SOTA) open-source MLLMs to generate high-quality video descriptions specifically tailored for BLV individuals?

To answer these questions, we conducted five user studies with both sighted and BLV individuals. Study 1 evaluated both open-source and proprietary MLLMs to determine the best model for

VideoA11y. A group of 150 sighted users on Amazon Mechanical Turk (MTurk) watched 150 videos from 15 categories (e.g., education, sports) and rated descriptions generated by VideoA11y using these MLLMs on four evaluation metrics: descriptiveness, objectivity, accuracy, and clarity. After selecting the most suitable model, the subsequent four studies focused on assessing the effectiveness of the VideoA11y method and the VideoA11y-40K dataset. In Study 2, another 150 sighted MTurk users watched the same 150 videos from Study 1 and rated descriptions generated by VideoA11y or novice humans on the same four metrics. In Study 3, 47 sighted MTurk users watched 47 YouTube videos and rated descriptions generated by VideoA11y and high-quality annotations produced by four members of our team, following the 42 AD guidelines. In Study 4, seven professional audio describers evaluated the same 47 videos and descriptions, selecting the better description for each video to further assess the quality of the annotations. Study 5 evaluated the alignment between video descriptions and the needs and satisfaction of BLV users. Specifically, 40 BLV users watched 10 videos from five categories with human-generated and VideoA11y descriptions. They rated each description based on the four metrics, selected their preferred description, and provided reasons for their choice. The results of these studies demonstrate that VideoA11y produces video descriptions of superior quality in all metrics compared to novice human annotations and is comparable to the quality of annotations produced by trained humans. Finally, we developed two complementary benchmarks to evaluate open-source MLLMs on VideoA11y and VideoA11y-40K, using standard NLP metrics and the four custom metrics of descriptiveness, objectivity, accuracy, and clarity. Our work is pioneering in the HCI and AI community, focusing on creating a video description dataset specifically for BLV

users and validating it with both sighted and BLV individuals. The novelty of this work lies in bridging established human practices of audio description with advancements in video description models and in creating a method, dataset, and benchmark dedicated to video accessibility. Our user studies and benchmark experiments demonstrate that MLLM-generated descriptions not only surpass the quality of novice human annotations but are also comparable to the standards of trained human annotations for video accessibility. In summary, the contributions of this paper are as follows:

- Develop VideoA11y, an MLLM-based approach for generating video descriptions using 42 AD guidelines that we collated to focus on the needs of BLV individuals.
- Release the first and most comprehensive video description dataset, VideoA11y-40K, for training models for BLV users.
- Demonstrate the effectiveness of VideoA11y and VideoA11y-40K via evaluation studies with 347 sighted participants, 40 BLV participants, and 7 professional audio describers.
- Introduce a new benchmark for video accessibility based on VideoA11y-40K.

2 Related Work

We review prior work on video accessibility, MLLMs for video understanding, and video description datasets and metrics.

2.1 Interactive Systems for Video Accessibility

Prior work on video accessibility aimed to simplify the task of sighted describers in writing descriptions [35, 38, 55]. LiveDescribe [8] developed an interface for novice volunteers to create audio descriptions. Similarly, Rescribe [64] assisted authors in creating and timing audio descriptions by optimizing content length and enabling iterative adjustments using dynamic programming. CrossA11y [47] further supports AD authors by detecting visual and auditory accessibility issues in videos, using cross-modal grounding analysis and an interface for reviewing and refining audio descriptions. However, these tools cannot generate partial or complete descriptions automatically. To address this issue, Yuksel et al. [85, 86] developed a human-in-the-loop machine learning approach in which the AI system provides initial video descriptions, and sighted novices edit the descriptions to enhance their quality. This approach improved the quality of video descriptions while reducing the time and effort required from volunteer describers. Yet, it still requires manual editing, making it difficult to scale. In response, Bodi et al. [7, 31] developed a fully automated system that generates descriptions and enables interactive question answering based on the visual content of a video.

More recent work used LLMs to summarize AI-generated descriptions for individual keyframes in a video. For example, Short-Scribe [71] leveraged automatic speech recognition (ASR), the image captioning model BLIP2 [42], and optical character recognition (OCR) to generate descriptions of several keyframes in a video, which are then summarized by GPT-4 to produce a video description. Similarly, SPICA [60] uses an image captioning model [73] to describe keyframes, followed by GPT-4 to turn the descriptions into a coherent narrative. These methods follow a hierarchical structure, generating descriptions for static frames first and then merging the descriptions with an LLM. This approach can result in missed context, inaccuracies during temporal changes, and semantic inconsistencies in the descriptions. In contrast, VideoA11y leverages MLLMs to process keyframes and generate video descriptions, preserving temporal information and minimizing semantic loss.

2.2 Multimodal Large Language Models for Video Understanding

Recent MLLMs demonstrate outstanding abilities in understanding, interpreting, and analyzing video content. MLLMs are trained on large multimodal (e.g., video, audio, text) datasets [16, 51, 87], then are fine-tuned or instruction-tuned for specific tasks. Fine-tuning involves taking a pre-trained model and training it further on a smaller, task-specific dataset. In video understanding, this could involve using a model pre-trained on general multimodal datasets and adapting it to tasks like video description or video question answering [16, 81, 84, 90]. Fine-tuning results in a highly specialized model for that task but may reduce its generalization capabilities across other tasks. On the other hand, instruction tuning enhances a model's ability to generalize across various tasks by improving how well it follows instructions [27, 43, 50, 76, 88]. The model is adjusted using diverse instructions that teach it how to interpret and perform different tasks. Our work leverages pre-trained MLLMs, which are subsequently fine-tuned on the proposed VideoA11y-40K dataset to benchmark their performance in generating accessible video descriptions for BLV users.

Other research has explored the use of prompt engineering to enhance model performance [9, 66, 77, 82]. Prompt engineering involves designing and optimizing inputs (prompts) to guide the model in generating relevant and accurate outputs without additional training on the pre-trained model. Previous work [9, 66] showed that both zero-shot and few-shot prompting can achieve performance comparable to fine-tuning without further training. Building on these studies [66], VideoA11y employs zero-shot prompt engineering to generate descriptions that adhere to AD guidelines and exceed the quality of human-generated descriptions.

2.3 Video Description Datasets

Numerous video description datasets have been introduced across various domains, including cooking [91], movies [3, 29, 70], social media [39], and human activities [13, 36, 45, 68, 69, 74]. Other datasets cover a broader range of video categories [4, 15, 17, 28, 75, 78, 79]. These datasets often include annotations from novice human describers recruited through online platforms, such as MTurk [13, 36, 39, 74, 79]. These human annotations can be brief, incomplete, and prone to spelling and grammar errors, especially when provided by inexperienced annotators [14, 33]. These issues affect the overall quality and usability of the dataset. VideoA11y addresses these issues by automatically generating accurate, grammatically correct descriptions from scratch while also correcting errors in human annotations found in existing video datasets and eliminating bias introduced by different annotators.

Recent datasets developed for video description have also used GPT as a part of their pipelines to aid in data generation [17, 50, 54, 75]. For example, VIDEOCC3M [54] leverages GPT-2 [66] as a decoder, utilizing features encoded by BERT [21] to generate video descriptions. InternVid [75] and Video-ChatGPT [50] apply

BLIP2 [42] and Tag2Text [30] to generate initial captions and synthesize them into video descriptions using Vicuna [65] or GPT-3.5. Meanwhile, Panda-70M [17] curates 3.8 million high-resolution videos from HD-VILA-100M [80] uses cross-modality teacher models for captioning, followed by fine-tuning a retrieval model on a selected subset to choose the best caption per video. Finally, OS-CaR [59] uses GPT-4V to create a dataset and benchmark for object state and state change description. While these approaches generate descriptions and provide benchmarks, they do not allow for tailoring the descriptions to the needs of BLV users, which is the focus of our work.

2.4 Evaluation Metrics for Video Descriptions

Evaluating video descriptions is essential for ensuring their quality and usability across various applications. Most video description evaluations rely on automated metrics, which are efficient, objective, and reproducible [1].

These metrics fall into two categories: n-gram-based and content-based. N-gram-based metrics like BLEU [63], METEOR [37], and CIDEr [72] measure n-gram overlap between generated and reference descriptions, focusing on precision and recall. Content-based metrics, such as SPICE [2], use scene graphs to compare objects, attributes, and relationships for semantic comparisons. However, these metrics often cannot fully capture the accessibility needs of BLV users.

Research involving BLV users frequently employs subjective evaluation methods. Yet, to our knowledge, no standard user-based metric exists for evaluating video descriptions. Existing work often asks participants to give an overall rating [71] for video descriptions or rate custom statements (e.g., "It provides me with useful additional information about the video") [60]. Recently, Natalie et al. created a qualitative codebook about different aspects of video description to guide novice describers in evaluating descriptions [57]. We build on this codebook by proposing four custom metrics tailored to BLV users' needs: descriptive, objective, accurate, and clear. These metrics provide a structured framework to guide human evaluators and ensure consistency in assessing video descriptions.

3 Overview of Our Process and Evaluation Metrics

Our Process. We developed and evaluated VideoA11y (method) and VideoA11y-40K (dataset) in four steps:

(1) **Developing the method: VideoA11y (Section 4):** We compiled 42 AD guidelines from online sources for professional describers and designed a compliant prompt based on these guidelines. To select an MLLM for VideoA11y, we applied our prompts to SOTA open-source model Video-LLaVA [43] and proprietary model GPT-4V [62] to create descriptions for 150 videos sampled from 15 different categories. Subsequently, we conducted a study with 150 sighted users on MTurk to evaluate the quality of the generated descriptions (Study 1).

(2) **Creating the dataset: VideoA11y-40K (Section 5):** Based on Step 1, we selected GPT-4V as the MLLM for VideoA11y to generate descriptions for 40,000 videos, resulting in the VideoA11y-40K dataset.

(3) **Evaluating VideoA11y and VideoA11y-40k with sighted novices, professional describers, and BLV users (Section 6):** We conducted four user studies to evaluate the method and dataset. The first two studies involved 197 sighted users: Study 2 compared the quality of video descriptions generated by our method with existing human annotations in video datasets, while Studies 3 and 4 asked sighted novices and professional describers to compare VideoA11y descriptions with high-quality annotations created by trained humans. Study 5 involved 40 BLV users who assessed the descriptions generated by VideoA11y compared to those produced by novice human describers.

(4) **Technical experiments to provide a benchmark for video accessibility (Section 7):** We fine-tuned two open-source MLLMs, Video-LLaVA [43] and LLaVA-Next-Video [89], on VideoA11y-40K. We then evaluated the fine-tuned models, along with baselines and other MLLMs, using a range of standard and custom metrics. This evaluation provides a benchmark for future video description models tailored to the needs of BLV users.

Metrics for Human and Technical Evaluations and Benchmarking. We employed a combination of custom and standard metrics to evaluate VideoA11y and VideoA11y-40K. For the custom metrics, we identified four specific metrics from the accessibility literature [57] and provided their definitions to participants in our studies. The four custom metrics are *descriptive*, *objective*, *accurate*, and *clear*. The *descriptive* metric evaluates whether the description provides detailed yet concise information about objects, people, and settings. The *objective* metric assesses whether only visible elements are reported without incorporating personal opinions or assumptions. The *accurate* metric focuses on the precision and correctness of details such as colors and spatial arrangements. The *clear* metric examines whether the information is presented in a way that is easy to follow and understand, avoiding confusion. Complete definitions of these metrics are provided in Appendix C. In addition, we used six standard metrics from the NLP domain: Bleu_1 [63], Bleu_4 [63], METEOR [37], ROUGE_L [44], CIDEr [72], SPICE [2]. These metrics assess various aspects of quality, including n-gram precision and recall (Bleu, METEOR, ROUGE) and semantic relevance and alignment with human judgment (CIDEr, SPICE).

4 Developing the Method: VideoA11y

VideoA11y employs MLLMs and video accessibility guidelines to produce precise and clear descriptions for BLV individuals. This process involves analyzing video frames and, when available, incorporating existing human annotations. We compiled AD guidelines from accessibility resources (Section 4.1), which were then integrated into a carefully crafted prompt. This compliant prompt, along with the keyframes (Section 4.2), is passed to an MLLM to generate or revise video descriptions (Section 4.3).

4.1 Curating Audio Description Guidelines

We initially collected a total of 154 AD guidelines from four different online sources: Netflix Accessibility Guidelines [58], Ofcom Guidelines [61], Media Access Canada Guidelines [52], and the Described and Captioned Media Program (DCMP) [20]. These guidelines are
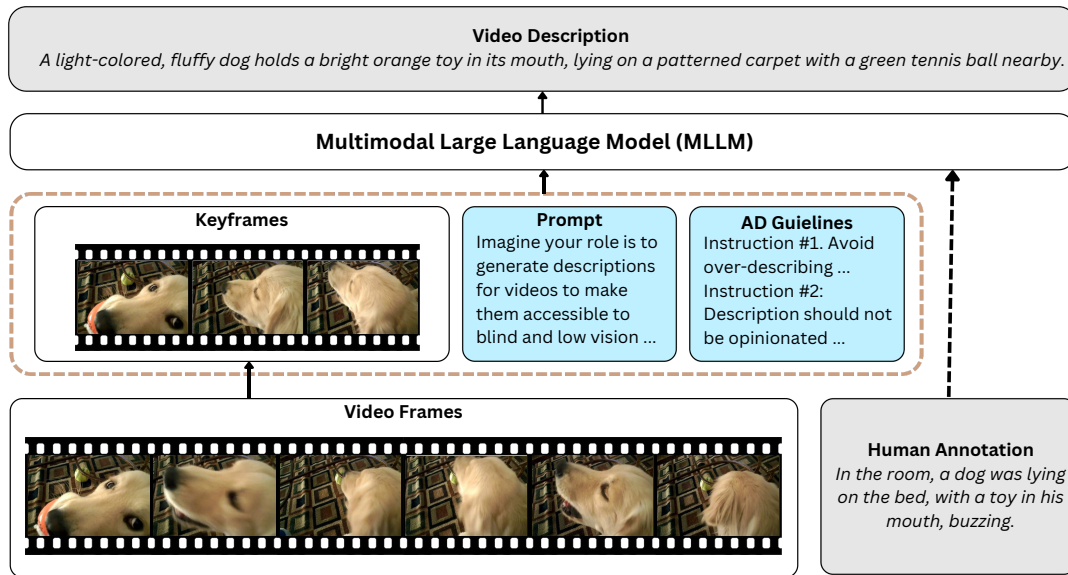
Figure 2: Overview of the VideoA11y pipeline. First, keyframes are extracted from the input video. Then, the keyframes, the prompt, AD guidelines, and optional human annotations are provided to MLLM, which generates accessible video descriptions.

curated for professional audio describers but have also been used to train novice describers to create descriptions for videos. They cover general guidelines for creating audio descriptions [61], as well as guidelines specific to educational [20] and entertainment content [52, 58]. While this list of 154 guidelines may not be exhaustive, they capture the majority of instructions agreed upon by professional describers. This was also evident from the overlap of several guidelines across these four resources. The overlapping and repeated guidelines were removed, and then the remainder were categorized based on whether an MLLM can be prompted to generate a description adhering to the guideline or not. This process removed guidelines focused on context, such as *"Description should include known relationships when they have been revealed."* [58], as well as those focused on voicing and audio of the video (e.g., *"Describe the source of sounds that may not be immediately recognizable within the video but are pertinent to understanding and appreciation of the content."*). Next, we shortened some of the guidelines for prompting. For example, *"Avoid over-describing — do not include visual images that are not vital to the understanding or enjoyment of the scene.",* became *"Avoid over-describing — Do not include non-essential visual details."* The result of this process was 42 AD guidelines optimized for prompting MLLMs (Appendix A).

4.2 Keyframe Extraction

Keyframes in a video capture significant changes or transitions within a scene, often representing shifts in content or visual focus. To extract keyframes from an input video, we implemented the local maximum algorithm [10, 23] which converts the frames from RGB to LUV color space to focus on luminance, then calculates the absolute difference between successive frames to measure the extent of change between frames. To reduce noise in the frame difference calculations, we applied a smoothing technique, which averages the values over a sliding window of 15 frames. This helps to smooth out minor fluctuations and highlight meaningful changes. Then, we identify peaks or local maxima, i.e., frames where the value is higher than the frames immediately before and after. These peaks represent significant changes in the video, indicating keyframes that likely correspond to scene transitions or important actions.

4.3 Prompt Design and Video Description Generation

We created a prompt using the AD guidelines:

> Imagine your role is to generate descriptions for videos to make them accessible to blind and low vision individuals. You will watch a sequence of keyframes from a video and read the current description of this video. Your task is to revise the current description. Output your result in a dictionary format: {"Video_Category": A string representing the category of video you believe it to be, "Revised_Desc": A string of revised description.}
>
> Current Description: {desc_current}
>
> Instructions:
> Instruction #1: Avoid over-describing — Do not include non-essential visual details.
> Instruction #2: Description should not be opinionated unless content demands it.
> Instruction #3: ...

We input the extracted keyframes and our compliant prompt, which includes the AD guidelines and optional human annotations, into an MLLM to generate or revise descriptions.
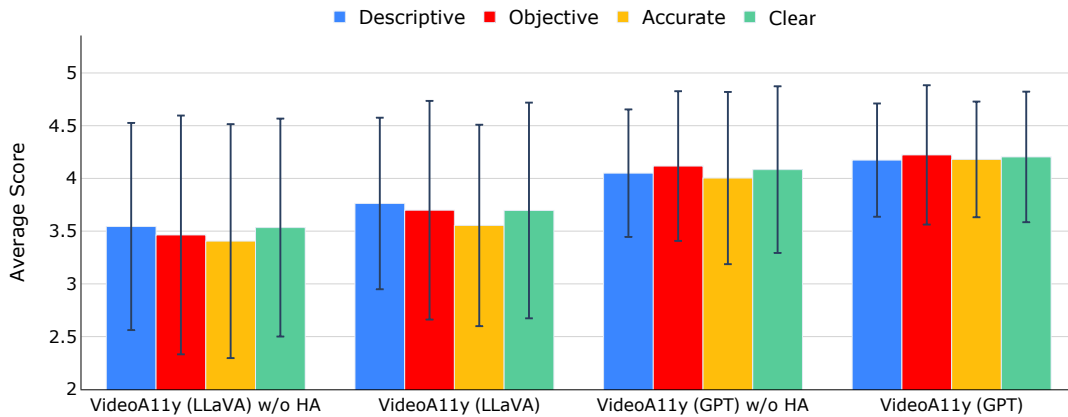
Figure 3: Results of Study 1 with 150 sighted MTurk users. VideoA11y (GPT) outperforms other methods on all metrics ($p < 0.05$), followed by VideoA11y (GPT) w/o HA. HA: Human Annotation.

4.4 Study 1: Evaluating VideoA11y on Open-Source and Proprietary MLLMs

We ran an MTurk experiment with sighted users to evaluate the difference in the quality of video descriptions generated by VideoA11y when using open-source vs. proprietary MLLMs. Specifically, we collected 150 videos sourced from three existing datasets: VALOR32K [15], VATEX [74], and YouCook2 [91] to run this evaluation. We used Video-LLaVA [43] as an open-source MLLM and GPT-4V [62] as a proprietary MLLM to generate descriptions, resulting in four conditions:

(1) **VideoA11y (LLaVA) w/o HA** uses the compliant prompt and Video-LLaVA to generate descriptions.
(2) **VideoA11y (LLaVA)** uses the compliant prompt with human annotations and Video-LLaVA to generate descriptions.
(3) **VideoA11y (GPT) w/o HA** uses the compliant prompt and GPT-4V to generate descriptions.
(4) **VideoA11y (GPT)** uses the compliant prompt with human annotations and GPT-4V to generate descriptions.

We recruited 150 MTurk participants for the study. The participants (74 males, 75 females, 1 preferred not to say) were between 23 and 60 years old and were primarily located in the United States. Each participant watched two videos and read four descriptions for each video. They then rated each description on a 5-point scale of extremely bad, somewhat bad, neither good nor bad, somewhat good, and extremely good for each of the four metrics. The user interface used in Study 1 is shown in Appendix D, and the full list of prompts for the four conditions is shown in Appendix B.

4.4.1 Study 1 Results. Figure 3 shows the average ratings for the four methods. We used the Friedman Test to analyze our data since the dependent variables (i.e., user ratings) are ordinal. The test reveals a significant effect of the description method. Pairwise comparisons (Appendix E.1) indicate that VideoA11y (GPT) w/o HA and VideoA11y (GPT) significantly outperform both VideoA11y (LLaVA) w/o HA and VideoA11y (LLaVA) in all four metrics ($p < 0.05$). The results also suggest that using human annotations as references can enhance the quality of descriptions, although not significantly ($p > 0.05$). Based on these results, we selected GPT-4V

as the MLLM and incorporated the existing human annotations in creating the VideoA11y-40K dataset.

5 Creating the Dataset: VideoA11y-40K

We employed VideoA11y (GPT) to generate high-quality video descriptions for three popular datasets in the computer vision community: VALOR32K [15] (29,635 videos), VATEX [74] (8,765 videos), and YouCook2 [91] (1,600 videos), in accordance with their MIT license permissions. This process resulted in the creation of the VideoA11y-40K dataset, which includes descriptions for 40,000 videos (32,000 training, 4,000 validation, and 4,000 test sets) across 15 categories tailored to BLV users.

5.1 Video Categorization

We derived 15 video categories by adapting and merging existing categories from YouTube to ensure comprehensive coverage. The categories are: (1) Film and Animation; (2) Music; (3) Sports; (4) Entertainment; (5) News and Politics; (6) Pets and Animals; (7) How-to and Instructional; (8) Event; (9) Travel, (10) People and Vlogs; (11) Food and Cooking; (12) Health and Wellness; (13) Auto and Technology, (14) Nonprofits and Activism; and (15) Education, Seminar and Talks. Each video in VideoA11y-40K was assigned to one of these 15 categories by using GPT-4 to analyze the video descriptions generated by VideoA11y. To verify categorization accuracy, we randomly sampled 5 videos from each category (75 videos in total) and recruited 225 MTurk participants (152 males, 71 females, aged 21–68 years) to rate the correctness of the assigned categories. A video was deemed misclassified if at least two out of three votes indicated an incorrect category. Our results confirmed that 96% (72 out of 75) of the videos were accurately categorized.

5.2 Dataset Statistics

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. Figure 4b presented the description length distribution in VideoA11y-40K, with the majority of captions ranging

Video Categories

(a) Distribution of video categories in the VideoA11y-40K dataset.

Video Description Length

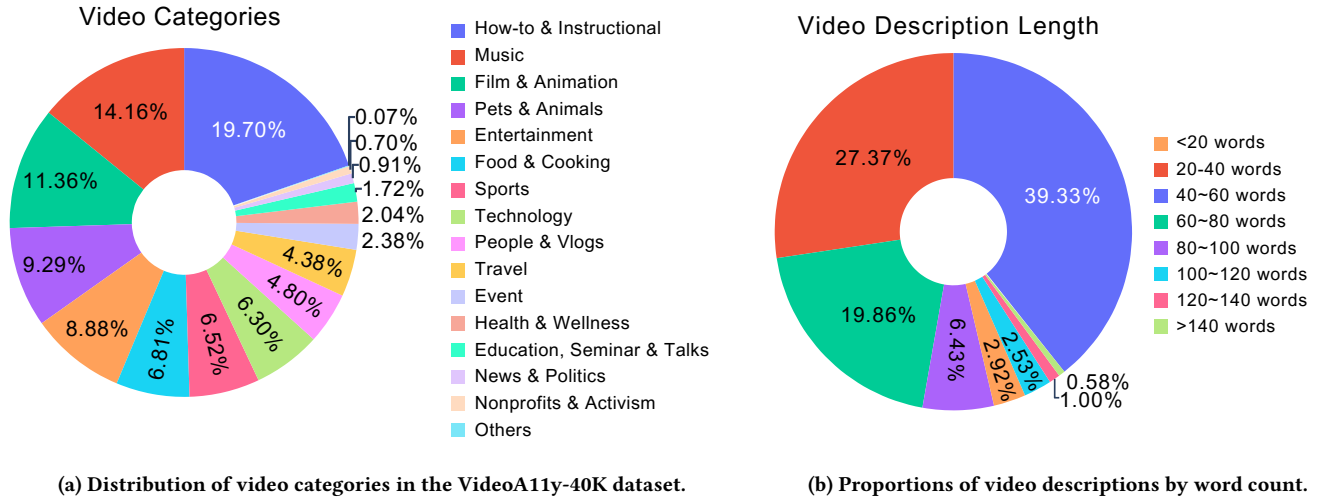(b) Proportions of video descriptions by word count.

Figure 4: Overview of video categories and description lengths in VideoA11y-40K.

between 40-60 words (39.33%) and 20-40 words (27.37%). Descriptions shorter than 20 words or longer than 140 words comprise only 2.92% and 0.58%, respectively. Figure 4a showed video distribution across 15 categories; How-to and Instructional (19.7%), Music (14.16%), and Film and Animation (11.36%) dominate, while News and Politics and Nonprofits and Activism each make up less than 1% of the videos.

6 Evaluating VideoA11y Method and VideoA11y-40K Dataset with Sighted and BLV Users

We evaluated VideoA11y and VideoA11y-40K through human subject studies with sighted and BLV individuals. Specifically, we sampled 150 videos from VideoA11y-40K, evenly distributed across the 15 categories introduced in Section 5.1 (10 videos per category). We conducted two studies with sighted users on MTurk (Sections 6.1, Section 6.2), one online study with professional describers (Section 6.3), and an online study with BLV participants (Section 6.4) to evaluate the quality of the video descriptions in the dataset. We obtained IRB approval for all studies. Participants viewed the informed consent sheet on the first page of the online surveys and provided consent by proceeding. For these studies, we set the statistical significance level at $p = 0.05$. The full statistical results are presented in Appendix E.

6.1 Study 2: Comparison with Descriptions Produced by Novice Humans

We ran an MTurk experiment with sighted users to assess the quality of the descriptions in VideoA11y-40K compared to descriptions created by novice annotators in the original datasets. We also included GPT-4V-generated descriptions that did not use the AD guidelines in the prompt (i.e., non-compliant prompt) to assess how they compare to novice human annotations. We recruited 150 new MTurk participants for the study. The participants (98 males and 52 females) were between 22 and 60 years old. Of these, 149 were

from the United States, and one was from Italy. Each participant watched two videos and rated the following five descriptions for each video:

(1) **Human Annotation** uses novice human annotations from the original datasets.
(2) **GPT-4V** uses the non-compliant prompt to generate descriptions.
(3) **GPT-4V w/ HA** uses the non-compliant prompt with human annotations to generate descriptions.
(4) **VideoA11y w/o HA** uses the compliant prompt to generate descriptions.
(5) **VideoA11y** uses the compliant prompt with human annotations to generate descriptions.

The user interface used in Study 2 and the list of prompts are shown in Appendices D and B, respectively.

6.1.1 Study 2 Results. Figure 5 shows the average ratings for the five methods. As in Study 1, we used the Friedman Test to analyze the ratings. The pairwise comparisons, with significance values for multiple comparisons (Appendix E.2), show that VideoA11y offers significant enhancements in all four metrics compared to Human Annotation, GPT-4V, and GPT-4V w/ HA, with all comparisons resulting in p-values under 0.001, suggesting the effectiveness of our approach in enhancing video description quality. In addition, VideoA11y w/o HA demonstrates statistically significant improvements over Human Annotation, GPT-4V, and GPT-4V w/ HA in all four metrics, with all p-values below 0.02. Finally, the overall performance of baseline GPT-4V and GPT-4V w/ HA is comparable to novice human annotations ($p > 0.05$) across all metrics, except for the 'descriptive' metric where baseline GPT-4V shows a significantly better performance with $p < 0.05$. These results indicate the effectiveness of AD guidelines in improving description quality beyond novice human annotations. While these results highlight the strengths of VideoA11y, minor inaccuracies can still occur in certain cases, as illustrated with examples in Appendix G.
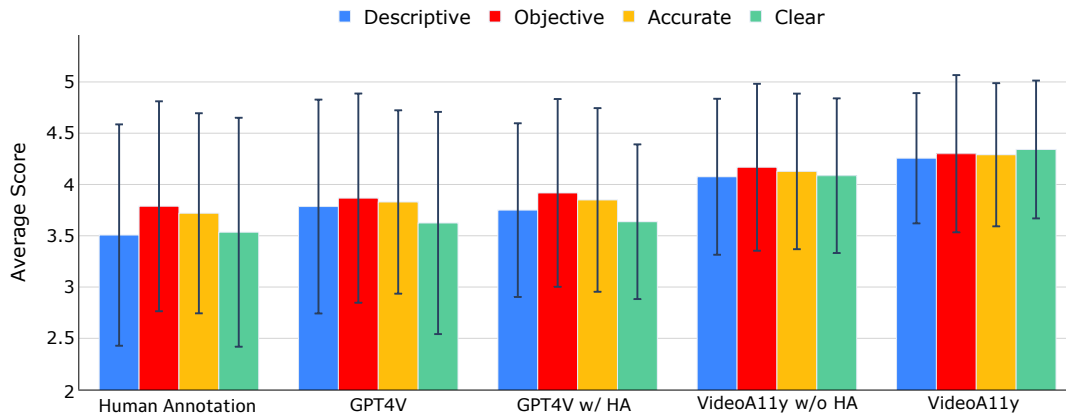
Figure 5: Results of Study 2 with 150 sighted MTurk users. VideoA11y outperforms other methods in all metrics ($p < 0.001$), followed by VideoA11y w/o HA. HA: Human Annotation.

6.2 Study 3: Comparison with Descriptions Produced by Trained Humans

In this study, we evaluated whether the descriptions generated by VideoA11y can meet the same standards as those produced by human describers who carefully follow AD guidelines. We selected 47 videos (ranging from 4 to 7 minutes long, with an average length of 4.92 minutes) from YouTube across various categories, and our team of four accessibility researchers created descriptions for these videos in accordance with the 42 AD guidelines we curated. We aimed to ensure that the description quality adheres to the standards and guidelines set by professional audio describers. We then recruited 47 sighted participants via MTurk to evaluate the descriptions generated from VideoA11y and those created by trained humans. In this study, we used GPT-4V as the MLLM for VideoA11y. On average, VideoA11y's descriptions (130.51 words) are comparable in length to human descriptions (140.17 words), ensuring a fair comparison. Figure 6a shows that the average ratings for VideoA11y descriptions are higher than those for high-quality human annotations in all four metrics (descriptiveness, objectivity, accuracy, and clarity). Furthermore, the Wilcoxon Signed-Rank test (Appendix E.3) demonstrates that the descriptions generated by VideoA11y show a statistically significant improvement ($p < 0.05$) in the 'clear' metric compared to high-quality human annotations. Thus, we used VideoA11y descriptions as the ground truth to conduct the additional technical experiments reported below.

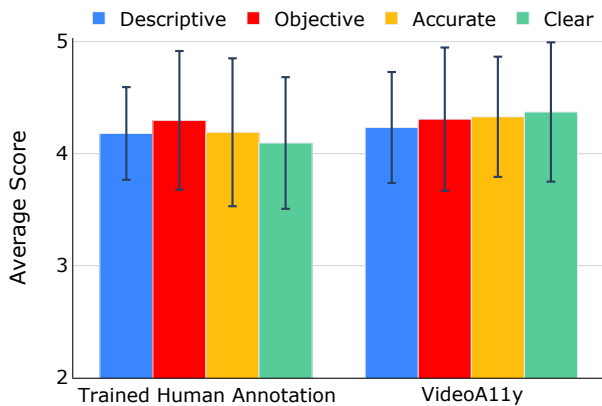## 6.3 Study 4: Evaluation with Professional Describers

We conducted a study with seven professional audio describers, each with over three years of paid experience (Appendix F.1), to evaluate the quality of VideoA11y's descriptions compared to those created by trained humans. Finding professional describers is challenging, and prior work on video accessibility included one to three professional describers [11, 34]. In our study, the seven experts viewed the 47 videos from Section 6.2, rated both sets of descriptions on four metrics, and selected the better description for each video with reasoning. Additionally, they participated in a Zoom

interview to provide qualitative insights. Figure 6b shows that the average expert ratings for VideoA11y descriptions exceed those for trained human annotations on all four metrics. Also, the differences in ratings between VideoA11y and trained humans are more noticeable when evaluated by expert describers. A Wilcoxon Signed-Rank test (Appendix E.4) reveals no statistically significant differences ($p > 0.05$) on all four metrics between VideoA11y and human descriptions, likely due to the small sample size. The medium to large effect sizes for three metrics (0.459–0.640) suggest the difference in the ratings is important. Additionally, professional audio describers preferred VideoA11y's descriptions for 33 (out of 47) videos.
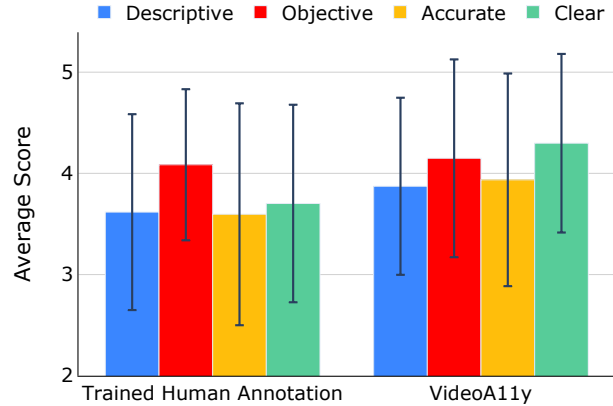
While unaware of the description source, the experts liked the choice of words, sentence structure, and visual details in VideoA11y's descriptions, noting these as important factors in conveying the visual feel and narrative intent of the videos ($n = 4$). They found VideoA11y's descriptions more accurate in describing the events ($n = 4$) and actions and described them as "engaging (P1)", "flavorful (P3)", and with good flow between sentences. Regarding visual detail, the experts noted that, overall, with a comparable length to human annotations, VideoA11y's descriptions included more information on character appearance, reactions, environment, and on-screen text compared to human descriptions. They suggested to include even more details on *"gender identity, race, approximate age, range, and body composition"* (P3).

In contrast, they stated the human annotations were objective, but in some cases "too literal (P5)" and "bland" (n=3): *"Sometimes it was the sentence structure, it was so boring and not artistic. I would choose the one that varied in pacing. (P1)"* Among the 14 videos with preferred human annotation, seven videos had notably longer descriptions than VideoA11y's. These descriptions included details on the physical appearance of people, scenes, and object features (e.g., a watch, types of rice). Also, humans better narrated the purpose of actions in two videos. These comments suggest that, in some cases, trained humans could surpass AI quality in providing details.

When asked if they could identify VideoA11y's descriptions from the human descriptions, participants could not distinguish between them ($n = 3$) or wrongly identified the human descriptions as AI-generated ($n = 4$). The experts thought the more objective and

(a) Results of Study 3 with 47 sighted MTurk users. VideoA11y received higher average ratings than trained humans on all metrics, with a statistically significant difference on the clear metric ($p = 0.004$).

(b) Results of Study 4 with seven professional describers. VideoA11y received higher average ratings than trained humans on all metrics, although the differences were not statistically significant ($p > 0.05$).

Figure 6: Comparison of VideoA11y's descriptions and trained human annotations on the 47 videos, evaluated by sighted MTurk users and professional describers.

literal descriptions were AI-generated (which was not the case). P5 justified: *"AI is really good at collecting information, but AI isn't great at integrating or resynthesizing it into a more human voice... Description 1 [VideoA11y's] tended to be more like how I would write them."* P4 and P7 made similar comments. This feedback highlights VideoA11y's ability to align more closely with established professional standards for description writing.

6.4 Study 5: Evaluation with Blind and Low Vision Individuals

We recruited 40 BLV participants to evaluate the effectiveness of VideoA11y and VideoA11y-40K. The participants' demographic information is shown in Appendix F.2. Six participants were completely blind, and 34 were legally blind with a visual acuity ranging from 20/200 to 20/1000 [19]. We selected 10 videos (2 per category) from (1) Entertainment, (2) How-to and Instructional, (3) Sports, (4) Pets and Animals, and (5) People and Vlogs in VideoA11y-40K dataset. We divided the participants into two groups of 20, with each group evaluating five videos. We inserted human annotations from the original datasets and VideoA11y-40K descriptions as audio descriptions at timestamps preceding the video segments they referred to according to the AD best practices [20, 52]. After reading the definition of the four evaluation metrics, BLV participants watched each video once with existing human descriptions and again with VideoA11y-40K descriptions with counterbalanced presentation order and rated the descriptions on four metrics of descriptiveness, objectiveness, accuracy, and clarity using a 10-point Likert scale. They also selected which description they preferred for each video and provided reasons for their selection without knowing the description source. We show the user interface in Appendix D.

6.4.1 Study 5 Results. Results from BLV users show that VideoA11y outperforms novice human describers in all five video categories and achieves a selection rate exceeding 80% in every category

(Figure 7a). Videoa11y had a selection rate of 90% (180 out of 200), demonstrating that our approach significantly enhanced the ability of BLV individuals to understand and enjoy video content. We also compared ratings for the four metrics between human annotations and VideoA11y (Figure 7b). Based on the Wilcoxon Signed-Rank test results in Table 1, VideoA11y significantly outperforms human annotations in all metrics, with ratings of 8.54 vs. 5.43 (descriptive), 8.33 vs. 5.79 (objective), 8.09 vs. 5.59 (accurate), and 8.36 vs. 5.29 (clear) with all p-values below 0.001. The results indicate VideoA11y's efficacy in enhancing video understanding for BLV users.

Comments from BLV users highlighted factors that impacted their description choices and ratings. 28 participants valued the clarity of descriptions generated by VideoA11y, while 25 highlighted the detailed descriptions facilitated their understanding of the videos and the context of the events. For example, P16 noted for the Entertainment video: *"At the end of the video, we had the man crying, the second audio [human annotated] failed to tell us why he was. The first audio [VideoA11y] however told us he did it 'playfully'. This is a very vital information as it adds a whole new context to it. The first audio [VideoA11y] also narrated the events accordingly."* In addition, 17 participants noted the accuracy and completeness of VideoA11y descriptions in providing details about specific actions and visual elements that were often missing or less comprehensive in the human annotations. P17 highlighted for the Pets and Animals video: *"It [VideoA11y] completely describes the actions of the woman from when she walked out with her horse and tied the horse. But the second video [human annotation] didn't talk about what is happening currently and at some points I was lost because I didn't understand what was being said."* P36 also mentioned this point for the Entertainment video: *"The second audio description [VideoA11y] was more detailed in term of object, colour, shape, clothing, the actions in the video, facial expressions of the man while the first [human annotated] seems a bit short of details."*
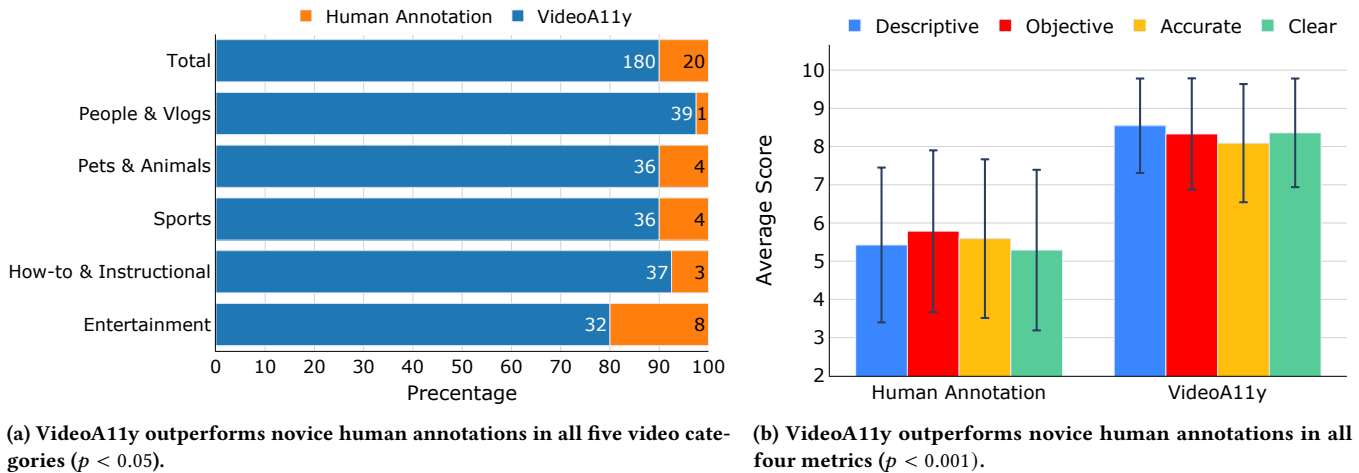
Chaoyu Li, Sid Padmanabhuni, Maryam Cheema, Hasti Seifi, and Pooyan Fazli

(a) VideoA11y outperforms novice human annotations in all five video categories ($p < 0.05$).

(b) VideoA11y outperforms novice human annotations in all four metrics ($p < 0.001$).

Figure 7: Results of Study 5 with 40 blind and low vision users.

Table 1: Overall pairwise comparisons from BLV user evaluations between VideoA11y and novice human descriptions in Study 5.

Condition 1	Condition 2	Metric	Effect Size	Test Statistics	P Value
VideoA11y	Human Annotaion	Descriptive	0.805	11.387	**<0.001**
VideoA11y	Human Annotaion	Objective	0.711	10.052	**<0.001**
VideoA11y	Human Annotaion	Accurate	0.708	10.014	**<0.001**
VideoA11y	Human Annotaion	Clear	0.772	10.920	**<0.001**

Each row tests the null hypothesis that the Condition 1 and Condition 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05. All effect sizes are large, indicating practical significance.

Furthermore, nine BLV participants noted that VideoA11y descriptions closely matched the unfolding events. For instance, P15 mentioned for the How-to and Instructional video: *"It [VideoA11y] described every action that took place accurately with the audio being in sync with the video."* Lastly, eight participants favored descriptions from VideoA11y for providing a more balanced and impartial depiction of events, ensuring that all aspects of the scene were given attention. P13 emphasized for the Sports video: *"In the first video [VideoA11y], the arm wrestling matches are described in more detail, and both men are assigned a shirt color so that viewers can place each contestant. In the second video [human annotation], the description focuses on the winner, and does not discuss the progression of the matches at all."* Overall, these responses underscore that video descriptions generated by VideoA11y enhanced the viewing experience for BLV users by offering clear, detailed, accurately synchronized, and less biased narratives compared to human annotations.

7 Technical Experiments to Provide a Benchmark for Video Accessibility

We benchmarked the performance of SOTA open-source models as candidates for the MLLM in VideoA11y through two complementary evaluations: (1) benchmarking the VideoA11y method

and (2) benchmarking models fine-tuned on the VideoA11y-40K dataset. Both evaluations used standard and custom metrics, with the VideoA11y-40K descriptions serving as the ground truth. This choice is supported by the studies above with sighted and BLV users, which show that the quality of VideoA11y-40K descriptions surpasses even that of trained human annotations. For benchmarking VideoA11y, we assessed its ability to enhance the performance of diverse MLLMs without fine-tuning to highlight its generalizability. For benchmarking VideoA11y-40K, we compared fine-tuned models against SOTA open-source baselines. All evaluations were conducted on the held-out test set of VideoA11y-40K.

7.1 Benchmarking VideoA11y

7.1.1 Baseline Models. We assessed four open-source models, including Video-LLaVA-7B [43], VILA1.5-40B [46], LLaVA-NeXT-Video-32B [89], and LLaVA-OneVision-72B [40]. To ensure a fair comparison, we adhered to the original settings for all models, including the number of frames and inference hyperparameters. We then used each MLLM to generate descriptions for the VideoA11y-40K test set without incorporating VideoA11y.

7.1.2 Results on Standard Metrics. Table 2 presents the results based on standard NLP metrics. Notably, all models show improvements across all metrics after integrating VideoA11y, with SPICE showing the largest increase. This suggests that VideoA11y enables models to generate descriptions that are semantically closer to the ground truth, even when the exact wording differs.

7.1.3 Results on Custom Metrics. We followed the method outlined in recent work [43, 49] to use GPT-4o as an evaluator for video descriptions. In this evaluation framework, GPT-4o treats VideoA11y-40K descriptions as the ground truth and assesses descriptions generated by other models on the four accessibility metrics, each rated on a scale from 1 to 5. Table 3 results indicate consistent improvements across all metrics and all evaluated models when applying VideoA11y. Overall, VideoA11y yields larger improvements in the accurate and clear metrics. Enhanced accuracy ensures that

Table 2: Comparison of standard NLP metrics for different models with and without VideoA11y on a held-out test set. Bold numbers indicate better performance for each model. +VA: w/ VideoA11y, -VA: w/o VideoA11y.

Model	Frames	Bleu_1		Bleu_4		METEOR		ROUGE_L		CIDEr		SPICE	
		-VA	+VA	-VA	+VA	-VA	+VA	-VA	+VA	-VA	+VA	-VA	+VA
Video-LLaVA-7B	8	12.28	**12.73**	1.08	**1.55**	12.99	**13.78**	16.06	**17.27**	0.63	**2.44**	11.09	**14.39**
VILA1.5-40B	8	18.96	**20.81**	4.08	**4.57**	12.55	**13.39**	20.15	**21.68**	8.87	**10.97**	17.72	**20.80**
LLaVA-NeXT-Video-32B	32	22.57	**24.48**	4.93	**5.34**	20.59	**21.99**	22.54	**23.50**	3.06	**3.21**	19.49	**22.14**
LLaVA-OneVision-72B	24	21.01	**32.25**	2.93	**7.01**	15.99	**17.73**	18.75	**23.50**	1.55	**14.59**	13.90	**21.32**

Table 3: Comparison of custom metrics for different models with and without VideoA11y on a held-out test set. Bold numbers indicate better performance for each model. +VA: w/ VideoA11y, -VA: w/o VideoA11y.

Model	Frames	Descriptive		Objective		Accurate		Clear	
		-VA	+VA	-VA	+VA	-VA	+VA	-VA	+VA
Video-LLaVA-7B	8	2.72	**2.89**	2.49	**2.64**	1.70	**2.10**	2.70	**2.91**
VILA1.5-40B	8	2.35	**2.38**	3.21	**3.48**	2.45	**2.52**	2.87	**3.02**
LLaVA-OneVision-72B	24	3.07	**3.18**	3.18	**3.46**	2.32	**2.70**	3.17	**3.49**
LLaVA-NeXT-Video-32B	32	3.68	**3.91**	3.34	**3.39**	2.76	**2.94**	3.67	**3.95**

Table 4: Comparison of standard NLP metrics for different models on a held-out test set. Bold number indicate the best performance, and <u>underlined</u> number indicate the second best performance.

Model	Frames	Bleu_1	Bleu_4	METEOR	ROUGE_L	CIDEr	SPICE
Video-LLaVA-7B	8	12.73	1.55	13.78	17.27	2.44	14.39
VILA1.5-40B	8	20.81	4.57	13.39	21.68	10.97	20.80
LLaVA-NeXT-Video-32B	32	24.48	5.34	<u>21.99</u>	23.50	3.21	22.14
LLaVA-OneVision-72B	24	32.25	7.01	17.73	23.50	14.59	21.32
VideoA11y-7B	8	<u>39.95</u>	<u>12.46</u>	21.11	<u>29.90</u>	<u>35.82</u>	26.98
VideoA11y-32B	32	**41.87**	**13.95**	**22.42**	**31.46**	**40.29**	**29.20**

Table 5: Comparison of custom metrics for different models on a held-out test set. Bold number indicate the best performance, and <u>underlined</u> number indicate the second best performance.

Model	Frames	Descriptive	Objective	Accurate	Clear
Video-LLaVA-7B	8	2.89	2.64	2.10	2.91
VILA1.5-40B	8	2.38	3.48	2.52	3.02
LLaVA-OneVision-72B	24	3.18	3.46	2.70	3.49
LLaVA-NeXT-Video-32B	32	<u>3.91</u>	3.39	2.94	<u>3.95</u>
VideoA11y-7B	8	3.45	<u>3.72</u>	<u>2.98</u>	3.82
VideoA11y-32B	32	**3.98**	**3.94**	**3.06**	**3.97**

descriptions faithfully capture the video content without including misleading or irrelevant information, while improved clarity provides well-structured and easily comprehensible descriptions. These improvements enable BLV users to better understand and engage with video content.

7.2 Benchmarking VideoA11y-40K

7.2.1 Baseline Models. We evaluated the same four open-source models described in Section 7.1.1 adhering to their original settings. VideoA11y then used each MLLM to generate descriptions for the VideoA11y-40K test set.

7.2.2 Fine-tuning Models on VideoA11y-40K. We fine-tuned Video-LLaVA-7B and LLaVA-NeXT-Video-32B on VideoA11y-40K. We refer to the fine-tuned models as VideoA11y-7B and VideoA11y-32B. For training, we employed Lora fine-tuning with a configuration

of rank 128 and alpha 256. Our fine-tuning parameters included 10 epochs, a learning rate of 2e-5, a batch size of 4 per device, and a maximum model length of 32,768. VideoA11y then used the fine-tuned models to generate descriptions for the videos in the VideoA11y-40K test set.

7.2.3 Results on Standard Metrics. We used our dataset to compute standard metrics for descriptions generated by baseline and fine-tuned models. Table 4 shows that VideoA11y-32B, followed by VideoA11y-7B, significantly outperforms other MLLMs in generating accessible video descriptions in all metrics. These results highlight VideoA11y-32B's ability to generate accurate, detailed, and semantically rich video descriptions, thereby enhancing video accessibility and user engagement.

7.2.4 Results on Custom Metrics. Table 5 shows the ratings provided by the GPT-4o evaluator, indicating that VideoA11y-32B

achieves the highest scores in all metrics. These results highlight the effectiveness of the VideoA11y-32B model in generating high-quality video descriptions for BLV users and confirm the value of the VideoA11y-40K dataset in improving video accessibility in machine learning models. The GPT-4o evaluator's prompt is shown in Appendix B.

8 Discussion

We present the answers to our research questions and discuss the value of VideoA11y and VideoA11y-40K for BLV users and video accessibility research. Moreover, we highlight the limitations of our work and outline future directions.

8.1 Reflection on Research Questions

We conducted five studies and four technical experiments to comprehensively assess the effectiveness of VideoA11y and VideoA11y-40K, focusing on answering three research questions: (1) How do VideoA11y descriptions compare in quality to those created by novice and trained human describers? (2) How do professional describers and BLV users evaluate and prefer VideoA11y descriptions compared to human descriptions? (3) Can the VideoA11y-40K dataset enhance SOTA open-source MLLMs to generate high-quality video descriptions specifically tailored for BLV individuals? Our study results consistently demonstrated the value of VideoA11y in addressing these questions. For RQ1, results from Study 2 showed that VideoA11y significantly outperformed novice human annotations in all metrics. In addition, in Study 3 and Study 4 results, VideoA11y descriptions were similar to high-quality descriptions produced by trained humans, as evidenced by evaluations from novice MTurk evaluators and professional describers. For RQ2, results from Study 4 and Study 5 demonstrated that both professional describers and BLV users preferred VideoA11y descriptions over human annotations. In Study 4, professional describers praised VideoA11y's narrative style and detailed visual representation, highlighting its alignment with professional standards. In Study 5, BLV users rated VideoA11y descriptions significantly higher than novice human annotations in all metrics and strongly preferred VideoA11y in over 90% of cases, enhancing their video comprehension and enjoyment. The BLV users' comments highlighted that VideoA11y descriptions had high clarity, matched with unfolding events, and were more balanced and impartial than human annotations. For RQ3, we compared competitive baseline models with two models fine-tuned on VideoA11y-40K and found that the models fine-tuned on VideoA11y-40K significantly outperformed the baselines in generating accessible video descriptions.

8.2 Implications of Our Method, Dataset, and Benchmark for Video Accessibility

VideoA11y enables supporting video accessibility for BLV users at scale. While hundreds of video description models have been introduced in the computer vision community over the past decade [1], none have addressed the real-world societal application of audio descriptions. This gap emphasized the importance and novelty of VideoA11y in accelerating progress in this underexplored area. Moreover, collecting high-quality human annotations for videos is hard to scale. When we collected high-quality human annotations

for 47 videos in Study 3 (Section 6.2), each annotator took an average of 3-4 minutes to describe 1 minute of a video. Even with meticulous effort, human annotations could not surpass VideoA11y's descriptions. In contrast, VideoA11y accelerates the creation of video descriptions, ensuring error-free outputs without grammatical mistakes, and is applicable to any video content, as demonstrated by our studies involving over 197 videos from 15 different categories. It can also be extended to incorporate additional guidelines and MLLMs that handle other modalities (e.g., audio). Thus, VideoA11y has the potential to be used on online video-sharing platforms (e.g., YouTube and TikTok) to create accessible content for BLV users.

VideoA11y produced minimal hallucinations. In the context of video description, hallucination refers to inaccuracies where the description includes details not present in the actual video content [41]. When VideoA11y was applied without human annotations as a reference (VideoA11y w/o HA), there were occasional instances where the model introduced actions or details not found in the video (see examples in Appendix G). However, when human annotations were incorporated as references, VideoA11y showed a reduction in hallucinations, as evidenced by the 'accuracy' ratings of over 4.2 out of 5 by 300 sighted users in Study 1 and Study 2. Additionally, we observed that BLV users had heightened sensitivity to the audio components of the videos, enabling them to detect inconsistencies between the audio content and the provided descriptions. Therefore, the high 'accuracy' scores by BLV users further demonstrate that VideoA11y descriptions are highly accurate.

VideoA11y can also shift the role of audio describers from generating descriptions to providing guidance and verification for MLLM-generated content. The effectiveness of VideoA11y reflects the richness and value of AD best practices and guidelines developed by professional describers over the past decades. In fact, several professional audio describers emphasized the importance of our four custom metrics—descriptive, objective, accurate, and clear—and noted that all are essential for evaluating and producing high-quality video descriptions. While larger datasets can improve MLLM output, our studies demonstrated that providing effective guidance to the model was equally important. Our results suggest a pathway for integrating MLLMs with the AD professional community to develop a human-AI collaboration [85] workforce for accessibility. For instance, in our experiments in Section 7.1, we demonstrated that applying the existing 42 guidelines improved MLLM performance across all metrics without any additional training. Building on this success, audio describers could further enhance the list of guidelines by adding new and revised rules for MLLMs, creating specific guidelines for different video genres (e.g., entertainment vs. people and vlogs, long- vs. short-form videos), and verifying the model's output. The model could also provide the list of guidelines used to generate a description, further facilitating the human verification and guidance process.

The VideoA11y-40K dataset, along with its benchmark, can support the development of future computer vision models. As the largest dataset on video accessibility, VideoA11y-40K enables researchers to train models across various video genres, and our benchmark can assess the robustness of these models. Our experiments in Section 7 demonstrated that fine-tuning open-source MLLMs on VideoA11y-40K led to significant improvements compared to baseline models in both standard and custom metrics.

While the performance of the fine-tuned models does not yet match that of the GPT-4V, they provide a more cost-effective and scalable solution. The scalability of fine-tuned open-source models is key to widespread adoption, enabling the continuous creation of tailored descriptions for BLV users at a significantly lower cost.

8.3 Limitations and Future Work

When human annotations are absent as references, relying solely on AD guidelines for VideoA11y to generate descriptions can lead to hallucinations. In our tests, these inaccuracies were mostly related to minor scene details. However, the broader impact of such inaccuracies in video descriptions on a larger scale remains largely unexplored. For example, during a public health crisis, inaccurate descriptions could prevent BLV individuals from receiving crucial information needed to make informed decisions. Additionally, VideoA11y is vulnerable to injection attacks [48], where human annotations might include false or harmful content, potentially spreading misinformation or negatively affecting the well-being of BLV users. Future research could explore the impact of these hallucinations across different content types and investigate ways to reduce inaccuracies. This may include applying direct preference optimization (DPO) during training to minimize hallucinations [67], using helper models for extracting specific information (e.g., object recognition [22], optical character recognition [26]) as input for prompts, or incorporating feedback from sighted volunteers to enhance the reliability of AI-generated descriptions.

Another limitation of VideoA11y is the lack of customization options tailored to the specific preferences and needs of BLV individuals. While our approach relies on general AD guidelines, it does not currently support personalized adjustments based on individual user preferences [25, 32, 56], such as preferred levels of detail or focus on specific types of content (e.g., action vs. emotional tone). Future research could gather more data on BLV users' preferences, such as the desired level of detail, focus of descriptions, and preferred style or tone across various video categories. These insights could enable VideoA11y to dynamically adapt to individual preferences, further enhancing comprehension and enjoyment for BLV users.

Lastly, VideoA11y currently lacks the ability to control the length and timing of descriptions to fit seamlessly within the natural pauses of a video, which is crucial for supporting inline descriptions. BLV users generally prefer inline descriptions that are integrated into the video flow without pausing or interrupting the experience [32]. To enable this, the system would need to dynamically adjust the length of descriptions and identify appropriate moments in the video to present them [64]. While these aspects were not the focus of our current work, future studies could integrate VideoA11y with systems like Rescribe [64] to support inline descriptions, enhancing the flexibility and usability of VideoA11y for BLV users and its integration into video-watching platforms.

9 Conclusion

This paper addresses a critical gap in video understanding research by developing a novel approach to enhance video accessibility for BLV individuals. We introduced VideoA11y, a method that leverages multimodal large language models and accessibility guidelines to generate video descriptions specifically tailored to the unique needs of BLV users. Rigorous experiments showed that VideoA11y outperformed both novice and trained human describers in terms of clarity, accuracy, objectivity, and descriptiveness, achieving higher satisfaction among BLV users. In addition, we presented the VideoA11y-40K dataset, the largest and most comprehensive video description dataset, comprising 40,000 videos described according to AD guidelines. Benchmarking VideoA11y-40K using both standard and custom metrics demonstrates that fine-tuned MLLMs on this dataset generate more accessible descriptions for BLV users, underscoring the potential of MLLMs to scale video content accessibility.

Acknowledgments

References

[1] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video Description: A Survey of Methods, Datasets, and Evaluation Metrics. *ACM Computing Surveys (CSUR)* 52, 6 (2019), 1–37.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision (ECCV)*.

[3] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed Movies: Story Based Retrieval with Contextual Embeddings. In *Asian Conference on Computer Vision (ACCV)*.

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE International Conference on Computer Vision (ICCV)*.

[5] Ava Bartolome and Shuo Niu. 2023. A Literature Review of Video-Sharing Platform Research in HCI. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[6] US Access Board. 2024. *About the ICT Accessibility 508 Standards and 255 Guidelines.* Retrieved Feb. 18, 2025 from https://www.access-board.gov/ict/

[7] 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. In *ACM SIGCHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*.

[8] Carmen J. Branje and Deborah I. Fels. 2012. LiveDescribe: Can Amateur Describers Create High-Quality Audio Description? *Journal of Visual Impairment & Blindness (JVIB)* 3 (2012), 154–165.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*. 1877–1901.

[10] Changqing Cao, Zehua Chen, Gang Xie, and Shaoshuai Lei. 2012. Key Frame Extraction Based on Frame Blocks Differential Accumulation. In *24th Chinese Control and Decision Conference (CCDC)*. 3621–3625.

[11] Ruei-Che Chang, Chao-Hsien Ting, Chia-Sheng Hung, Wan-Chen Lee, Liang-Jin Chen, Yu-Tzu Chao, Bing-Yu Chen, and Anhong Guo. 2022. OmniScribe: Authoring Immersive Audio Descriptions for 360° Videos. In *ACM Symposium on User Interface Software and Technology (UIST)*.

[12] Maryam Cheema, Hasti Seifi, and Pooyan Fazli. 2024. Describe Now: User-Driven Audio Description for Blind and Low Vision Individuals. arXiv:2411.11835 [cs.HC]

[13] David L. Chen and William B. Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

[14] Haoran Chen, Jianmin Li, Simone Frintrop, and Xiaolin Hu. 2022. The MSR-Video to Text Dataset with Clean Annotations. *Computer Vision and Image Understanding (CVIU)* (2022), 103581.

[15] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023. VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset. arXiv:2304.08345 [cs.LG].

[16] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[17] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Cheng-Yu Chuang and Pooyan Fazli. 2023. CLearViD: Curriculum Learning for Video Description. arXiv:2311.04480 [cs.CV].

[19] Lalit Dandona and Rakhi Dandona. 2006. Revision of visual impairment definitions in the International Statistical Classification of Diseases. *BMC Medicine* 4 (2006).

[20] DCMP. 2024. *Description Key - Quality Description*. Retrieved Feb. 18, 2025 from https://dcmp.org/learn/621-description-key---quality-description

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186.

[22] Yongqi Ding, Lin Zuo, Mengmeng Jing, Pei He, and Yongjun Xiao. 2024. Shrinking Your TimeStep: Towards Low-Latency Neuromorphic Object Recognition with Spiking Neural Networks. In *AAAI Conference on Artificial Intelligence (AAAI)*. 11811–11819.

[23] Jingjin Du, Yale Zhao, Shanna Zhuang, and Zhengyou Wang. 2021. Key Frame Extraction for Falling Detection. In *International Conference on Information Technology and Biomedical Engineering (ICITBE)*. 105–109.

[24] Federal Communications Commission (FCC). 2024. *Twenty-First Century Communications and Video Accessibility Act*. Retrieved Feb. 18, 2025 from https://www.fcc.gov/cvaa

[25] Nazaret Fresno, Judit Castellà, and Olga Soler-Vilageliu. 2016. *'What Should I Say?' Tentative Criteria to Prioritize Information in the Audio Description of Film Characters*. Palgrave Macmillan UK, 143–167.

[26] Masato Fujitake. 2024. DTrOCR: Decoder-Only Transformer for Optical Character Recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 8025–8035.

[27] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2023. VTimeLLM: Empower LLM to Grasp Video Moments. arXiv:2311.18445 [cs.CV].

[28] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020. Multimodal Pretraining for Dense Video Captioning. In *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*. 470–490.

[29] Q. Huang, Y. Xiong, A. Rao, J. Wang, and D. Lin. 2020. MovieNet: A Holistic Dataset for Movie Understanding. In *European Conference on Computer Vision (ECCV)*. 709–727.

[30] Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. 2024. Tag2Text: Guiding Vision-Language Model via Image Tagging. In *The Twelfth International Conference on Learning Representations (ICLR)*.

[31] Shasta Ihorn, Yue-Ting Siu, Aditya Bodi, Lothar Narins, Jose M Castanon, Yash Kant, Abhishek Das, Ilmi Yoon, and Pooyan Fazli. 2021. NarrationBot and InfoBot: A Hybrid System for Automated Video Description. arXiv:2111.03994 [cs.CV].

[32] Lucy Jiang, Crescentia Jung, Mahika Phutane, Abigale Stangl, and Shiri Azenkot. 2024. "It's Kind of Context Dependent": Understanding Blind and Low Vision People's Video Accessibility Preferences Across Viewing Scenarios. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[33] Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. *Computational Linguistics* (2023), 157–198.

[34] Masato Kobayashi, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa. 2009. Providing Synthesized Audio Description for Online Videos. In *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. 249–250.

[35] Masatomo Kobayashi, Trisha O'Connell, Bryan Gould, Hironobu Takagi, and Chieko Asakawa. 2010. Are Synthesized Video Descriptions Acceptable?. In *International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. pp. 163–170.

[36] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *International Conference on Computer Vision (ICCV)*.

[37] Alon Lavie and Abhaya Agarwal. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Association for Computational Linguistics (ACL)*. 228–231.

[38] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In *Association for Computational Linguistics (ACL)*. 2603–2614.

[39] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval. In *European Conference on Computer Vision (ECCV)*. 447–463.

[40] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326 [cs.CV].

[41] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. 2024. VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding. arXiv:2412.03735 [cs.CV].

[42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *40th International Conference on Machine Learning (ICML)*.

[43] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. arXiv:2311.10122 [cs.CV].

[44] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Association for Computational Linguistics (ACL)*. 74–81.

[45] Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. VideoXum: Cross-modal Visual and Textural Summarization of Videos. *IEEE Transactions on Multimedia (TMM)* (2023).

[46] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2024. VILA: On Pre-training for Visual Language Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[47] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang Anthony Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding. In *ACM Symposium on User Interface Software and Technology (UIST)*.

[48] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt Injection attack against LLM-integrated Applications. arXiv:2306.05499 [cs.CV].

[49] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. MMBench: Is Your Multi-modal Model an All-around Player?. In *European Conference on Computer Vision (ECCV)*.

[50] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[51] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *International Conference on Computer Vision (ICCV)*.

[52] B. Milligan and D. Fels. 2012. *Media Access Canada (MAC) - Our Projects - Descriptive Video Production and Presentation Best Practices Guide for Digital Environments*. Retrieved Feb. 18, 2025 from http://www.mediac.ca/DVBPGDE_V2_28Feb2012.asp

[53] Valerie S. Morash, Yue-Ting Siu, Joshua A. Miele, Lucia Hasty, and Steven Landau. 2015. Guiding Novice Web Workers in Making Image Descriptions Using Templates. *ACM Transactions on Accessible Computing (TACCESS)* (2015).

[54] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. 2022. Learning Audio-Video Modalities from Image Captions. In *European Conference on Computer Vision (ECCV)*. 407–426.

[55] Rosiana Natalie. 2022. Cost-effective and Collaborative Methods to Author Video's Scene Description for Blind People.. In *ACM SIGCHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*.

[56] Rosiana Natalie, Ruei-Che Chang, Smitha Sheshadri, and Kotaro Hara Anhong Guo. 2024. Audio Description Customization. In *International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. 1–19.

[57] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-Ren Chan, Ebrima H Jarjue, Hernisa Kacorri, and Kotaro Hara. 2021. The Efficacy of Collaborative Authoring of Video Scene Descriptions. In *ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*.

[58] Netflix. 2024. *Audio Description Style Guide v2.5*. Retrieved Feb. 18, 2025 from https://partnerhelp.netflixstudios.com/hc/en-us/articles/215510667-Audio-Description-Style-Guide-v2-5

[59] Nguyen Nguyen, Jing Bi, Ali Vosoughi, Yapeng Tian, Pooyan Fazli, and Chenliang Xu. 2024. OSCaR: Object State Captioning and State Change Representation. In *In Findings of the Association for Computational Linguistics (NAACL)*.

[60] Zheng Ning, Brianna L Wimer, Kaiwen Jiang, Keyi Chen, Jerrick Ban, Yapeng Tian, Yuhang Zhao, and Toby Jia-Jun Li. 2024. SPICA: Interactive Video Content Exploration through Augmented Audio Descriptions for Blind or Low-Vision Viewers. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[61] Ofcom. 2021. *Ofcom's Guidelines on the Provision of Television Access Services*. Retrieved Feb. 18, 2025 from https://www.ofcom.org.uk/__data/assets/pdf_file/0025/212776/provision-of-tv-access-services-guidelines.pdf

[62] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL].

[63] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Association for Computational Linguistics (ACL)*. 311–318.

[64] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *ACM Symposium on User Interface Software and Technology (UIST)*. 747–759.

[65] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. arXiv:2304.03277 [cs.CV].

[66] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. *Language Models are Unsupervised Multitask Learners*. Technical Report 1. OpenAI.

[67] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[68] G.A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *European Conference on Computer Vision (ECCV)*. 510–526.

[69] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 2018. Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos. arXiv:1804.09626 [cs.CV].

[70] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. MAD: A Scalable Dataset for Language Grounding in Videos From Movie Audio Descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5026–5035.

[71] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making Short-Form Videos Accessible with Hierarchical Video Summaries. In *ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[72] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4566–4575.

[73] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *International Conference on Machine Learning (ICML)*.

[74] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *International Conference on Computer Vision (ICCV)*.

[75] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. In *International Conference on Learning Representations (ICLR)*.

[76] Zhanyu Wang, Longyue Wang, Minghao Wu, Zhen Zhao, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. 2023. GPT4Video: A Unified Multimodal Large Language Model for Instruction-Followed Understanding and Safety-Aware Generation. *Computing Research Repository (CoRR)* (2023).

[77] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[78] Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, Chenliang Li, Qi Qian, Maofei Que, Ji Zhang, Xiao Zeng, and Fei Huang. 2023. Youku-mPLUG: A 10 Million Large-scale Chinese Video-Language Dataset for Pre-training and Benchmarks. arXiv:2306.04362 [cs.CV].

[79] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5288–5296.

[80] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

[81] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[82] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large Language Models as Optimizers. In *International Conference on Learning Representations (ICLR)*.

[83] YouDescribe. 2024. *YouDescribe*. Retrieved Feb. 18, 2025 from https://www.youdescribe.org/

[84] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Wenbing Tao. 2023. Merlin: Empowering Multimodal LLMs with Foresight Minds. arXiv:2312.00589 [cs.CV].

[85] Beste Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Seung Jung Jin, Yue-Ting Siu, Joshua A Miele, and Ilmi Yoon. 2020. Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users. In *ACM Conference on Designing Interactive Systems (DIS)*. 47–60.

[86] Beste F Yuksel, Soo Jung Kim, Seung Jung Jin, Joshua Junhee Lee, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Ilmi Yoon, Yue-Ting Siu, and Joshua A Miele. 2020. Increasing video accessibility for visually impaired users with human-in-the-loop machine learning. In *ACM SIGCHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*. 1–9.

[87] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[88] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*. 543–553.

[89] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. *LLaVA-NeXT: A Strong Zero-shot Video Understanding Model*. Retrieved Feb. 18, 2025 from https://llava-vl.github.io/blog/2024-04-30-llava-next-video/

[90] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning Video Representations from Large Language Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[91] Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards Automatic Learning of Procedures from Web Instructional Videos. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Chaoyu Li, Sid Padmanabhuni, Maryam Cheema, Hasti Seifi, and Pooyan Fazli

A   Audio Descriptions Guidelines

The list below shows the complete 42 audio description (AD) guidelines we curated for VideoA11y.

Instruction #1. Avoid over-describing — Do not include non-essential visual details.

Instruction #2. Description should not be opinionated unless content demands it.

Instruction #3. Choose level of detail based on plot relevance when describing scenes.

Instruction #4. Description should be informative and conversational, in present tense and third-person omniscient.

Instruction #5. The vocabulary should reflect the predominant language/accent of the program and should be consistent with the genre and tone of the content while also mindful of the target audience. Vocabulary used should ensure accuracy, clarity, and conciseness.

Instruction #6. Consider historical context and avoid words with negative connotations or bias.

Instruction #7. Pay attention to verbs — Choose vivid verbs over bland ones with adverbs.

Instruction #8. Use pronouns only when clear whom they refer to.

Instruction #9. Use comparisons for shapes and sizes with familiar and globally relevant objects.

Instruction #10. Maintain consistency in word choice, character qualities, and visual elements for all audio descriptions.

Instruction #11. Tone and vocabulary should match the target audience's age range.

Instruction #12. Ensure no errors in word selection, pronunciation, diction, or enunciation.

Instruction #13. Start with general context, then add details.

Instruction #14. Describe shape, size, texture, or color as appropriate to the content.

Instruction #15. Use first-person narrative for engagement if required to engage the audience.

Instruction #16. Use articles appropriately to introduce or refer to subjects.

Instruction #17. Prefer formal speech over colloquialisms, except where appropriate.

Instruction #18. When introducing new terms, objects, or actions, label them first, and then follow with the definitions.

Instruction #19. Describe objectively without personal interpretation or comment. Also, do not censor content.

Instruction #20. Deliver narration steadily and impersonally (but not monotonously), matching the program's tone.

Instruction #21. It can be important to add emotion, excitement, lightness of touch at different points. Adjust style for emotion and mood according to the program's genre.

Instruction #22. If it is children's content, tailor language and pace for children's TV, considering audience feedback.

Instruction #23. Do not alter, filter, or exclude content. You should describe what you see. Try to seek simplicity and succinctness in your description.

Instruction #24. Prioritize what is relevant when describing action as to not affect user experience.

Instruction #25. Include location, time, and weather conditions when relevant to the scene or plot.

Instruction #26. Focus on key content for learning and enjoyment when creating audio descriptions. This is so that the intention of the program is conveyed.

Instruction #27. When describing an instructional video/content, describe the sequence of activities first.

Instruction #28. For a dramatic production, include elements such as style, setting, focus, period, dress, facial features, objects, and aesthetics.

Instruction #29. Describe what is most essential for the viewer to know in order to follow, understand, and appreciate the intended learning outcomes of the video/content.

Instruction #30. The description should describe characters, locations, on-screen action, and on-screen information.

Instruction #31. Describe only what a sighted viewer can see.

Instruction #32. Describe main and key supporting characters' visual aspects relevant to identity and personality. Prioritize factual descriptions of traits like hair, skin, eyes, build, height, age, and visible disabilities. Ensure consistency and avoid singling out characters for specific traits. Use person-first language.

Instruction #33. If unable to confirm or if not established in the plot, do not guess or assume racial, ethnic or gender identity.

Instruction #34. When naming characters for the first time, aim to include a descriptor before the name (e.g., "a bearded man, Jack").

Instruction #35. Description should convey facial expressions, body language and reactions.

Instruction #36. When important to the meaning / intent of content, describe race using currently-accepted terminology.

Instruction #37. Avoid identifying characters solely by gender expression unless it offers unique insights not apparent otherwise to low vision viewers.

Instruction #38. Describe character clothing if it enhances characterization, plot, setting, or genre enjoyment.

Instruction #39. If text on the screen is central to understanding, establish a pattern of on-screen words being read. This may include making an announcement, such as "Words appear".

Instruction #40. In the case of subtitles, the describer should read the translation after stating that a subtitle appears.

Instruction #41. When shot changes are critical to the understanding of the scene, indicate them by describing where the action is or where characters are present in the new shot.

Instruction #42. Provide description before the content rather than after.

B   Prompts and Implementation Details

B.1   Prompt for GPT-4V

The following prompt was employed for the *GPT-4V* method in Study 2, as detailed in Section 6.1.

> Imagine your role is to generate descriptions for videos. You will watch a sequence of keyframes from a video and craft a description based on these keyframes.

B.2 Prompt for GPT-4V w/ HA

The following prompt was employed for the *GPT-4V w/ HA* method in Study 2, as detailed in Section 6.1.

> Imagine your role is to generate descriptions for videos. You will watch a sequence of keyframes from a video and read the current description of this video. Your task is to revise the description.

B.3 Prompt for VideoA11y w/o HA

The following prompt was employed for the *VideoA11y (LLaVA) w/o HA* and the *VideoA11y (GPT) w/o HA* method in Study 1 (as outlined in Section 4.4), and *VideoA11y w/o HA* method in Study 2 (Section 6.1).

> Imagine your role is to generate descriptions for videos to make them accessible to blind and low vision individuals. You will watch a sequence of keyframes from a video. Based on these keyframes, craft a description. You must follow all the given instructions. You should avoid any prefatory language, such as 'the video shows'. Output your result as a dictionary format: {"Video_Category": A string representing the category of video you believe it to be, "Revised_Desc": A string of description.}

B.4 Prompt for VideoA11y

The following prompt was employed for the *VideoA11y (LLaVA)* and the *VideoA11y (GPT)* methods in Study 1 (as outlined in Section 4.4), and the VideoA11y method in Study 2 (Section 6.1).

> Imagine your role is to generate descriptions for videos to make them accessible to blind and low vision individuals. You will watch a sequence of keyframes from a video and read the current description of this video. Your task is to revise the current description. You must follow all the given instructions. Output your result in a dictionary format: {"Video_Category": A string representing the category of video you believe it to be, "Revised_Desc": A string of revised description.}

B.5 Prompt for GPT-4o Evaluator

The following prompt was employed for the GPT-4o evaluator used in technical experiments (as outlined in Section 7.2.4).

> You are an expert evaluator with a deep understanding of video description quality, particularly for making content accessible to blind and low vision (BLV) individuals. Your role is to assess and rate video descriptions based on specific metrics.

> Task:
> I have two video descriptions: one is the ground truth, and the other is generated by a model. Additionally, I have four evaluation metrics: Descriptive, Objective, Accurate, and Clear. Please evaluate the model-generated description using the above metrics. Provide a rating from 1 to 5 for each metric, and briefly explain the reasons for each rating.
>
> Metrics Definition:
> 1. Descriptive: A descriptive description should provide vivid details about objects, people, and settings while maintaining a concise narrative flow. It should include essential information about the appearance of individuals, such as their clothing, facial expressions, and actions, and visual properties of objects, such as color and shape. For example, "A smiling woman, wearing a loose white dress, types on a laptop in a softly lit room."
> 2. Objective: An objective description should report what is visible without adding personal opinions or assumptions. For instance, describe two people as "a woman and a man" without adding any relationship context unless it is mentioned. It should also avoid guessing personal attributes like racial or gender identities unless explicitly clear.
> 3. Accurate: An accurate description should aim for precision in describing visible elements without imagination. It should ensure that all details, such as colors and spatial arrangements, are reported correctly. For instance, "Blue sky with white clouds" instead of "White sky with blue clouds" if that is what appears. Additionally, it should avoid adding unnecessary details that do not contribute to a deeper understanding of the scene.
> 4. Clear: A clear description should present information in the videos in a way that is easy to follow for blind and low vision individuals. It should describe the object or character's properties before the actions or relationships with other objects or characters. For example, "A man wearing sunglasses plays the piano." Pronouns should only be used when it is clear who they refer to, and the description should include any on-screen text if it is central to understanding. For instance, "A man in a black polo shirt is speaking. He is in a studio setting with a red couch and a blue background featuring the text 'Talk of the Town'".
>
> Input:
> - Ground truth video description: {desc_gt}
> - Model-generated video description: {desc_can}
>
> Output Format:
> Return the result as a string format dictionary with the following structure:
> {"Descriptive": [Rating out of 5],
> "Objective": [Rating out of 5],
> "Accurate": [Rating out of 5],
> "Clear": [Rating out of 5],
> "Reason": "Your brief explanation here"}

Chaoyu Li, Sid Padmanabhuni, Maryam Cheema, Hasti Seifi, and Pooyan Fazli

Table 6: Evaluation metrics used in the user studies. Sighted MTurk and BLV participants reviewed these definitions before rating the video descriptions.

Metric	Description
Descriptive	A descriptive description should provide vivid details about objects, people, and settings while maintaining a concise narrative flow. It should include essential information about the appearance of individuals, such as their clothing, facial expressions, and actions, and visual properties of objects, such as color and shape. For example, "A smiling woman, wearing a loose white dress, types on a laptop in a softly lit room."
Objective	An objective description should report what is visible without adding personal opinions or assumptions. For instance, describe two people as "a woman and a man" without adding any relationship context unless it is mentioned. It should also avoid guessing personal attributes like racial or gender identities unless explicitly clear.
Accurate	An accurate description should aim for precision in describing visible elements without imagination. It should ensure that all details, such as colors and spatial arrangements, are reported correctly. For instance, "Blue sky with white clouds" instead of "White sky with blue clouds" if that is what appears. Additionally, it should avoid adding unnecessary details that do not contribute to a deeper understanding of the scene.
Clear	A clear description should present information in the videos in a way that is easy to follow for blind and low vision individuals. It should describe the object or character's properties before the actions or relationships with other objects or characters. For example, "A man wearing sunglasses plays the piano." Pronouns should only be used when it is clear who they refer to, and the description should include any on-screen text if it is central to understanding. For instance, "A man in a black polo shirt is speaking. He is in a studio setting with a red couch and a blue background featuring the text 'Talk of the Town'".

C   Metrics Definition

Table 6 provides a comprehensive definition of our four customized metrics for Study 1 (Section 4.4), Study 2 (Section 6.1), Study 3 (Section 6.2), Study 4 (Section 6.3), and Study 5 (Section 6.4). These definitions were also presented to participants in all studies (See Appendix D).

D   User Study Interfaces

D.1   User Interface of Studies 1, 2 and 3

Figure 8 illustrates the user interface used in Studies 1, 2 and 3. After providing informed consent on the first page of the online survey, participants watched a video, followed by reading the definitions of the four metrics proposed in Section 3. Subsequently, they were presented with multiple video descriptions generated by different methods (four in Study 1, five in Study 2, and two in Study 3). Each description was rated from "Extremely bad" to "Extremely good" based on the aforementioned metrics. To ensure fairness, all video descriptions were presented in a randomized order. This same procedure was then repeated with a long video.

D.2   User Interface of Study 5

Figure 9 illustrates the user interface used in Study 5. After providing informed consent on the first page of the online survey, BLV participants read the definitions of four metrics proposed in Section 3. They were then presented with a video paired with a human-annotated video description and they rated the quality of the description on a scale from 1 to 10, for each of the four metrics. This extended rating scale, from 1 to 10, was adopted following feedback from BLV users during pilot testing, who indicated that a 5-point scale was inadequate for BLV individuals to capture the nuanced variations in video descriptions. Subsequently, participants watched the same video accompanied by the VideoA11y-40K description, and again rated it using the same set of metrics. The sequence in which the human-annotated and VideoA11y-40K descriptions were presented was randomized to mitigate order bias. After evaluating both descriptions, participants were asked to rank them and provide justifications for their preferences. This process was replicated across four additional videos to ensure robust assessment.

D.3   User Interface of Video Category Evaluation Study

Figure 10 illustrates the user interface for the video category evaluation study. After providing informed consent on the first page of the online survey, participants viewed a video that had been pre-categorized by VideoA11y. They were then tasked with verifying the appropriateness of the assigned category on page 10a. If participants deemed the category incorrect ("False"), they were redirected to page 10b, where they could reassign the video to one of the 15 video categories.

D.4   User Interface of Demographic Questionnaire

Figure 11 shows the demographic questionnaire used across all studies. The questionnaire was designed to collect data on participants' age, gender, ethnicity, race, and country of residence. Details of the demographic data are presented in the results section of each study in Section 6.

E   Statistical Analysis of the Studies

E.1   Study 1: Statistical Analysis

Table 7 provides additional statistical analysis for Study 1 (Section 4.4). We applied the related-samples Friedman test, followed by post hoc pairwise comparisons, to analyze the data. Statistically significant results ($p < 0.05$) are highlighted in bold. These results demonstrate that VideoA11y (GPT) and VideoA11y (GPT) w/o HA show statistically significant improvements over VideoA11y (LLaVA) and VideoA11y (LLaVA) w/o HA in all four metrics.

E.2   Study 2: Statistical Analysis

Table 8 provides additional statistical analysis for Study 2 (Section 6.1). The data is analyzed using the related-samples Friedman test, followed by post hoc pairwise comparisons. Significant results ($p < 0.05$) are highlighted in bold. The analysis reveals that VideoA11y and VideoA11y w/o HA show statistically significant superiority in all four metrics when compared to Human Annotation, GPT-4V, and GPT-4V w/ HA.

Table 7: Overall pairwise comparisons evaluating VideoA11y on open-source and proprietary MLLMs in Study 1. HA: Human Annotation.

Condition 1 | Condition 2	Metric	Test Statistics	P Value
VideoA11y (GPT) w/o HA | VideoA11y (LLaVA) w/o HA	Descriptive	4.048	**<0.001**
	Objective	5.708	**<0.001**
	Accurate	5.313	**<0.001**
	Clear	4.095	**<0.001**
VideoA11y (GPT) w/o HA | VideoA11y (LLaVA)	Descriptive	2.546	**0.011**
	Objective	4.048	**<0.001**
	Accurate	4.411	**<0.001**
	Clear	3.178	**0.001**
VideoA11y (GPT) | VideoA11y (LLaVA) w/o HA	Descriptive	5.455	**<0.001**
	Objective	6.293	**<0.001**
	Accurate	6.625	**<0.001**
	Clear	5.550	**<0.001**
VideoA11y (GPT) | VideoA11y (LLaVA)	Descriptive	3.953	**<0.001**
	Objective	4.633	**<0.001**
	Accurate	5.724	**<0.001**
	Clear	4.633	**<0.001**
VideoA11y(GPT) | VideoA11y (GPT) w/o HA	Descriptive	1.407	0.159
	Objective	0.585	0.559
	Accurate	1.312	0.189
	Clear	1.455	0.146

Each row tests the null hypothesis that the Condition 1 and Condition 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

E.3 Study 3: Statistical Analysis

Table 9 provides additional statistical analysis for Study 3 (Section 6.2). We applied a Wilcoxon Signed-Rank test to compare the performance of VideoA11y vs. high-quality Human Annotations for the four metrics. Significant results ($p < 0.05$) are highlighted in bold. The analysis reveals that VideoA11y shows statistically significant superiority on the clear metric compared to Human Annotation.

E.4 Study 4: Statistical Analysis

Table 10 provides additional statistical analysis for Study 4 (Section 6.3). We applied a Wilcoxon Signed-Rank test to compare the performance of VideoA11y vs. high-quality Human Annotations for the four metrics. The analysis reveals that there are no statistically significant differences ($p > 0.05$) in all four metrics between VideoA11y and human descriptions, likely due to the small sample size. The medium to large effect sizes for three metrics (0.459–0.640) suggest the difference in the ratings is practically important.

F Demographic Information of Participants

F.1 Demographic Information of Professional Audio Describers

Table 11 shows the demographic information of seven professional audio describes in Study 4 (Section 6.3).

F.2 Demographic Information of BLV Participants

Table 12 shows the demographic information of 40 BLV participants in Study 5 (Section 6.4). The participants include 28 males, 11 females, and 1 individual who prefers not to specify their gender.

The age range spans from 18 to 51 years old. The majority are from the United States (39 participants), with one participant from the United Kingdom.

G Qualitative Results

Figure 12 and 13 illustrate qualitative examples where we compare Human Annotations with the descriptions generated by VideoA11y.

Figure 14 illustrates examples of hallucination phenomena observed in the descriptions generated by VideoA11y w/o HA. This figure provides a comparative analysis of descriptions from Human Annotators, VideoA11y w/o HA, and VideoA11y. In the absence of human annotation as a reference, GPT-4V sometimes introduces actions or details that are not present in the video or provide incorrect information. For instance, for the first video, the model described actions such as "sorts through" and "placing envelopes through a door's mail slot", which are not in the video content. Furthermore, for the last video, the movements of Tai Chi were incorrectly classified as a "dance routine" without the hint from human annotations.

Figure 15 illustrates examples of minor inaccuracies in descriptions generated by VideoA11y. For the first example, the phrase "against a smoky backdrop" is a hallucination, as no smoky backdrop is present in the actual video. Additionally, the description "a large illuminated cross as the centerpiece" is somewhat misleading, as the cross is part of the stage's backdrop rather than being the central focus of the scene. In the second example, the man does not "put on" and "take off" magnifying eyewear but merely gestures with it. Furthermore, the statement "He wears magnifying eyewear while using the saw, which is clamped to a table" is incorrect, as the copper sheet—not the saw—is clamped to the table.

Please watch the video below and then complete the rating scale for each of the five descriptions.

Evaluation Metrics
Please read the de initions o the our metrics carefully be ore you begin reading the video descriptions.

Descriptive	A descriptive description should provide vivid details about objects, people, and settings while maintaining a concise narrative flow. It should include essential information about the appearance of individuals, such as their clothing, facial expressions, and actions, and visual properties of objects, such as color and shape. For example, "A smiling woman, wearing a loose white dress, types on a laptop in a softly lit room."
Objective	An objective description should report what is visible without adding personal opinions or assumptions. For instance, describe two people as "a woman and a man" without adding any relationship context unless it is mentioned. It should also avoid guessing personal attributes like racial or gender identities unless explicitly clear.
Accurate	An accurate description should aim for precision in describing visible elements without imagination. It should ensure that all details, such as colors and spatial arrangements, are reported correctly. For instance, "Blue sky with white clouds" instead of "White sky with blue clouds" if that is what appears. Additionally, it should avoid adding unnecessary details that do not contribute to a deeper understanding of the scene.
Clear	A clear description should present information in the videos in a way that is easy to follow for blind and low vision individuals. It should describe the object or character's properties before the actions or relationships with other objects or characters. For example, "A man wearing sunglasses plays the piano." Pronouns should only be used when it is clear who they refer to, and the description should include any on-screen text if it is central to understanding. For instance, "A man in a black polo shirt is speaking. He is in a studio setting with a red couch and a blue background featuring the text 'Talk of the Town'".

Note: A longer description does not always mean a better quality.

Description #1:

A person in a blue t-shirt with 'Nineteen 75' on the back stands in a crowd. In the foreground, a child in a polka dot dress with a pink sash sits on the grass. The crowd is gathered at an outdoor event with a live band performing on stage.

	Extremely bad	Somewhat bad	Neither good nor bad	Somewhat good	Extremely good
Descriptive	○	○	○	○	○
Objective	○	○	○	○	○
Accurate	○	○	○	○	○
Clear	○	○	○	○	○

Description #2:

A child in a blue 'Nineteen 75' t-shirt and plaid shorts stands watching a performance, with people seated on the grass around them.
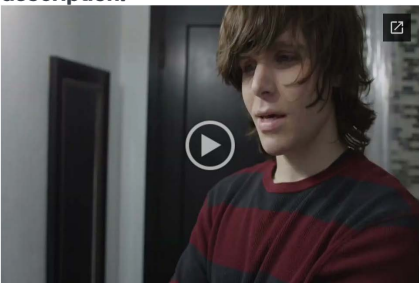
	Extremely bad	Somewhat bad	Neither good nor bad	Somewhat good	Extremely good
Descriptive	○	○	○	○	○
Objective	○	○	○	○	○
Accurate	○	○	○	○	○
Clear	○	○	○	○	○

Figure 8: User interface in Studies 1, 2 and 3. MTurk participants in Study 1 rated four video descriptions from different methods, MTurk participants in Study 2 rated five video descriptions generated by different methods. MTurk participants in Study 3 rated two video descriptions generated by VideoA11y and human. Video descriptions were presented to participants in random order.

Evaluation Metrics
Please read the definitions of the four metrics carefully before you begin watching the videos.

1. **Descriptive**: A descriptive description should provide vivid details about objects, people, and settings while maintaining a concise narrative flow. It should include essential information about the appearance of individuals, such as their clothing, actions, facial expressions, and visual properties of objects, such as color and shape. For example, "A smiling woman, wearing a loose white dress, types on a laptop in a softly lit room."
2. **Objective**: An objective description should report what is visible without adding personal opinions or assumptions. It should avoid guessing personal attributes like racial or gender identities unless explicitly clear. For instance, describe two people as "a woman and a man" without adding any relationship context unless it is mentioned.
3. **Accurate**: An accurate description should aim for precision in describing visible elements without imagination. It should ensure that all details, like colors and spatial arrangements, are reported correctly. For instance, "Blue sky with white clouds" instead of "White sky with blue clouds" if that is what appears. Additionally, it should avoid adding unnecessary details that do not contribute to a deeper understanding of the scene.
4. **Clear**: A clear description should present information in the videos in a way that is easy to follow for blind and low vision people. It should describe the object or character's properties before the actions or relationships with other objects or characters. For example, "A man wearing sunglasses and a baseball cap walks around a neighborhood." Pronouns should only be used when it is clear who they refer to, and the description should include any on-screen text if it is central to understanding. For instance, "A man in a black polo shirt is speaking. He is in a studio setting with a red couch and a blue background featuring the text 'Talk of the Town'".

Please watch the video below and then complete the rating scale for the audio description.

Please rate the audio description in this video from 1 to 10 (1 indicates extremely bad and 10 indicates extremely good) based on the "Descriptive" metric. (A descriptive description should provide vivid details about objects, people, and settings while maintaining a concise narrative flow. It should include essential information about the appearance of individuals, such as their clothing, actions, facial expressions, and visual properties of objects, such as color and shape. For example, "A smiling woman, wearing a loose white dress, types on a laptop in a softly lit room.")

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

Please watch the video again with another audio description and then complete the rating scale for the audio description.

Please rate the audio description in this video from 1 to 10(1 indicates extremely bad and 10 indicates extremely good) based on the "Descriptive" metric. (A descriptive description should provide vivid details about objects, people, and settings while maintaining a concise narrative flow. It should include essential information about the appearance of individuals, such as their clothing, actions, facial expressions, and visual properties of objects, such as color and shape. For example, "A smiling woman, wearing a loose white dress, types on a laptop in a softly lit room.")

1	2	3	4	5	6	7	8	9	10
○	○	○	○	○	○	○	○	○	○

Please choose the better audio description from the two audio descriptions for this video.

○ First Aduio Description
○ Second Audio Description

Please provide a reason to explain your ranking.

Figure 9: User interface in Study 5. We carefully designed all content to be fully accessible to BLV users via screen reader. This figure displays only the rating question for descriptiveness for both video descriptions due to space constraints. In the actual study, participants rated each video description against all four metrics. After evaluating two descriptions for the same video, participants ranked them and justified their rankings.

Please watch the video below and answer the following questions below.

This video falls under the Technology category.

○ True
○ False

(a) The interface for collecting responses to determine whether the video category is correct.

If you answered **False** for the previous question, categorize the video into one of the following categories below to the best of your abilties.

○ Sports
○ Entertainment
○ Pets & Animals
○ How-to & Instructional
○ Music
○ Nonprofits & Activism
○ Film & Animation
○ Event
○ Travel
○ Education, Seminar & Talks
○ News & Politics
○ Technology
○ Food & Cooking
○ People & Blogs
○ Health & Wellness

(b) The interface for collecting responses to re-categorize the video.

Figure 10: User interface of the category evaluation study for VideoA11y-40K. (a) MTurk participants watched a video and the category assigned by VideoA11y first and then determined whether the video category was correct. (b) If they selected "False," they were redirected to this page to assign a new category to the video.

Figure 11: The interface of the demographic questionnaire for all human studies.

Chaoyu Li, Sid Padmanabhuni, Maryam Cheema, Hasti Seifi, and Pooyan Fazli

Table 8: Overall pairwise comparisons between VideoA11y and other methods in Study 2. HA: Human Annotation.

Condition 1 | Condition 2	Metric	Test Statistics	P Value
GPT-4V | Human Annotation	Descriptive	3.033	**0.002**
	Objective	1.129	0.259
	Accurate	0.738	0.461
	Clear	1.381	0.167
GPT-4V w/ HA | Human Annotation	Descriptive	1.760	0.078
	Objective	1.313	0.189
	Accurate	0.962	0.336
	Clear	0.079	0.937
VideoA11y w/o HA | Human Annotation	Descriptive	5.515	**<0.001**
	Objective	4.228	**<0.001**
	Accurate	4.401	**<0.001**
	Clear	5.090	**<0.001**
VideoA11y w/o HA | GPT-4V	Descriptive	2.482	**0.013**
	Objective	3.099	**0.002**
	Accurate	3.663	**<0.001**
	Clear	3.709	**<0.001**
VideoA11y w/o HA | GPT-4V w/ HA	Descriptive	3.755	**<0.001**
	Objective	2.915	**0.004**
	Accurate	3.439	**<0.001**
	Clear	5.011	**<0.001**
VideoA11y | Human Annotation	Descriptive	7.156	**<0.001**
	Objective	6.066	**<0.001**
	Accurate	6.483	**<0.001**
	Clear	8.116	**<0.001**
VideoA11y | GPT-4V	Descriptive	4.123	**<0.001**
	Objective	4.937	**<0.001**
	Accurate	5.745	**<0.001**
	Clear	6.735	**<0.001**
VideoA11y | GPT-4V w/ HA	Descriptive	5.397	**<0.001**
	Objective	4.753	**<0.001**
	Accurate	5.521	**<0.001**
	Clear	8.037	**<0.001**

Each row tests the null hypothesis that the Condition 1 and Condition 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

Table 9: Overall pairwise comparisons between VideoA11y and trained human descriptions in Study 3.

Condition 1 | Condition 2	Metric	Test Statistics	P Value
VideoA11y | Human Annotation	Descriptive	0.551	0.582
VideoA11y | Human Annotation	Objective	0.238	0.812
VideoA11y | Human Annotation	Accurate	1.191	0.234
VideoA11y | Human Annotation	Clear	2.843	**0.004**

Each row tests the null hypothesis that the Condition 1 and Condition 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

Table 10: Overall pairwise comparisons between VideoA11y and trained human descriptions in Study 4.

Condition 1 | Condition 2	Metric	Effect Size	Test Statistics	P Value
VideoA11y | Human Annotation	Descriptive	0.459	1.214	0.225
VideoA11y | Human Annotation	Objective	0.288	0.762	0.446
VideoA11y | Human Annotation	Accurate	0.515	1.363	0.173
VideoA11y | Human Annotation	Clear	0.640	1.693	0.090

Each row tests the null hypothesis that the Condition 1 and Condition 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is 0.05.

Table 11: Participant demographics of professional audio describes in Study 4.

Pseudonym	Age	Gender	Ethnicity	Race	Country	Years of Experience
P1	50	Female	Not Hispanic or Latino	White	United States	7
P2	32	Female	Not Hispanic or Latino	Black	United States	5
P3	33	Female	Not Hispanic or Latino	White	United States	4
P4	30	Male	Not Hispanic or Latino	White	Canada	3
P5	26	Non-Binary	Not Hispanic or Latino	White	United States	4
P6	66	Female	Not Hispanic or Latino	White	United States	23
P7	30	Male	Not Hispanic or Latino	Black	United States	5

Table 12: Participant demographics in the BLV study. The description of vision is self-reported by participants.

Pseudonym	Age	Gender	Ethnicity	Race	Country	BLV Level
P1	27	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/800)
P2	28	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/1000)
P3	28	Male	Not Hispanic or Latino	White	United States	Legally Blind (20/500 - 20/1000)
P4	26	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/700)
P5	28	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/500)
P6	27	Male	Not Hispanic or Latino	White	United States	Legally Blind (20/500)
P7	44	Male	Not Hispanic or Latino	Asian	United States	Totally Blind
P8	33	Unknown	Hispanic or Latino	More than one race	United States	Totally Blind
P9	32	Female	Hispanic or Latino	Unknown	United States	Totally blind
P10	23	Female	Hispanic or Latino	Black	United States	Legally Blind (20/1000)
P11	26	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/500)
P12	24	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/500)
P13	29	Male	Not Hispanic or Latino	White	United States	Legally Blind (20/500 - 20/1000)
P14	26	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/900)
P15	22	Female	Hispanic or Latino	Unknown	United States	Totally blind
P16	31	Male	Not Hispanic or Latino	White	United States	Legally Blind (20/700)
P17	26	Male	Not Hispanic or Latino	Black	United States	Totally Blind
P18	23	Female	Not Hispanic or Latino	White	United States	Legally Blind (20/200)
P19	20	Male	Not Hispanic or Latino	Unknown	United States	Legally Blind (20/400)
P20	21	Female	Hispanic or Latino	Black	United States	Legally Blind (20/1000)
P21	28	Female	Not Hispanic or Latino	Black	United States	Legally Blind (20/600)
P22	25	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/200)
P23	23	Male	Not Hispanic or Latino	White	United States	Legally Blind (20/200)
P24	27	Female	Not Hispanic or Latino	White	United States	Legally Blind (20/200)
P25	28	Male	Not Hispanic or Latino	White	United States	Legally Blind (20/200)
P26	21	Male	Hispanic or Latino	White	United States	Legally Blind (20/500)
P27	24	Male	Not Hispanic or Latino	White	United States	Legally Blind (20/600)
P28	20	Female	Hispanic or Latino	White	United States	Legally Blind (20/500)
P29	21	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/400)
P30	21	Female	Not Hispanic or Latino	Black	United States	Legally Blind (20/500)
P31	22	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/400)
P32	21	Male	Hispanic or Latino	White	United States	Legally Blind (20/500)
P33	20	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/500)
P34	Unknown	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/400)
P35	30	Male	Hispanic or Latino	White	United States	Legally Blind (20/400)
P36	27	Female	Not Hispanic or Latino	Black	United States	Legally Blind (20/400)
P37	31	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/500)
P38	27	Male	Not Hispanic or Latino	White	United Kingdom	Legally Blind (20/1000)
P39	18	Male	Not Hispanic or Latino	Black	United States	Legally Blind (20/500)
P40	51	Female	Not Hispanic or Latino	White	United States	Totally Blind

Human Annotation: A mouse clicks on a screen with several lines of English and then moves into the game as a man speaks, with <u>several cartoon characters jumping up</u>.

VideoA11y: The video showcases gameplay from a *sandbox video game*. The player *navigates through a menu, selecting 'Play Multiplayer' and entering a server IP address*. *After clicking 'Connect', the screen displays 'Logging in...'* before transitioning to the game environment. The player is in a virtual world with a *vast blue sky* and *floating blocks representing different materials*. The player **moves around**, **placing and breaking blocks**, while *text chat from other players appears on the screen*.

Human Annotation: prepare some chicken cutlets and immense them in two beat eggs. coat the chicken cutlets with mixture of grated parmesan cheese and panko breadcrumbs. fry the chicken cutlets on a pan with oil on both sides. pour some marinara sauce on top of the fried chicken. place a piece of cheese on each piece. put the chicken pieces in an oven for about 10 minutes until cheese to melt.

VideoA11y: *A woman in purple shirt* shows a cooking tutorial of chicken parmesan. The process begins with chicken cutlets being dipped in beaten eggs, then coated with a mixture of grated parmesan cheese and panko breadcrumbs. The woman frys the coated cutlets in oil until golden brown on both sides. After frying, she spoons marinara sauce over each cutlet, followed by a slice of cheese. The chicken is then baked in a *400-degree Fahrenheit* oven until the cheese melts and becomes bubbly. *The final dish is served with a side of zucchini noodles.*

Figure 12: Qualitative comparison between human annotations and descriptions generated by VideoA11y. <u>Red underline</u> indicates the errors in human annotations, green bold indicates the corrected facts, and *blue italics* indicates additional details.

Human Annotation: An older man and older woman show how to do some sort of <u>meditation move</u>.

VideoA11y: Two individuals are in *a spacious room with wooden floors*. The man *on the left wears a black t-shirt, black pants, and white sneakers*. The woman *on the right wears a blue t-shirt, black pants, and white sneakers*. There is *a yin-yang logo on their t-shirts*. They demonstrate a **Tai Chi sequence** with a *focus on footwork*.

Human Annotation: Many men screamed in pain when they fell while skating on the field.

VideoA11y: Skateboarders attempt challenging tricks at various locations. One skater *loses balance while landing a stair jump, tumbling onto the pavement*. Another skater *slips off a rail, falling onto their back*. A third skater *stumbles during an indoor ramp trick, sliding down the ramp*.

Human Annotation: Letters popped up on the screen, then a man stood on a blue stage and spoke to the music.

VideoA11y: The video opens with a *blue and white animated logo with the text 'Howcast'* transitioning into text that reads *'5 Advanced Bouldering Techniques for Indoor Climbing'*. A man with a *beard, wearing a blue t-shirt labeled 'The Cliffs'*, introduces himself in a spacious indoor climbing gym with *colorful climbing holds on the walls*. He stands on a blue padded floor with *a chalk bag beside him*, ready to demonstrate climbing techniques.

Figure 13: Qualitative comparison between human annotations and descriptions generated by VideoA11y. <u>**Red underline**</u> **indicates the errors in human annotations,** **green bold** **indicates the corrected facts, and** *blue italics* **indicates additional details.**

Chaoyu Li, Sid Padmanabhuni, Maryam Cheema, Hasti Seifi, and Pooyan Fazli

Human Annotation: the mail woman is leaving a note for somebody at their house

VideoA11y w/o HA: A hand <u>sorts through</u> a large stack of envelopes. A postal worker *in a light blue shirt and grey shorts delivers mail to a house*, <u>placing envelopes through a door's mail slot</u>. The worker is then seen writing on a <u>clipboard</u>. Finally, *a postal worker in a similar uniform speaks with a reporter beside a delivery vehicle*.

VideoA11y: A hand **holds** a stack of envelopes. A postal worker *in a light blue shirt and navy shorts knocks on the door at a residence* and then **leaves a notice in pen on an envelope**. *A colleague in a similar uniform speaks to a reporter beside a delivery vehicle*.

Human Annotation: A woman in a short white sleeve plays a white instrument and makes a wonderful sound. A woman in a red dress sings.

VideoA11y w/o HA: Two women perform *on a city sidewalk*. One, in *traditional Bavarian dress*, <u>dances and claps</u> while the other, in *casual attire*, *plays an accordion*.

VideoA11y: *On a city sidewalk*, two women perform. One, *dressed in a traditional outfit with a white blouse, red skirt, and apron*, **sings**. The other, in a *casual white top and blue jeans, plays an accordion*. They *stand before a metal barrier*, and *passersby can be seen in the background*.

Figure 14: Qualitative comparison between human annotations, VideoA11y w/o HA, and VideoA11y. HA: Human Annotation. <u>Red underline</u> indicates the hallucinations from VideoA11y w/o HA, green bold indicates the corrected facts, and *blue italics* denotes correct additional details that are absent in the human annotations.

VideoA11y: The video opens with the title 'Women Worship Warriors 2021' in ornate script against a smoky backdrop. A live music performance unfolds on a stage bathed in blue and purple lights, with a large illuminated cross as the centerpiece. A lead female singer in a sparkling blue dress holds the audience's attention with her powerful vocals, using a handheld microphone. She is accompanied by a band consisting of a drummer, a saxophonist, a guitarist, and a keyboard player, each deeply engaged with their instruments. Backing vocalists in coordinated blue outfits provide harmonies, adding to the rich musical tapestry. The audience, visible in the foreground, watches intently.

VideoA11y: A man with glasses and a gray beard, wearing a Rio Grande shirt, sits in a workshop filled with tools and equipment. His name is Mark Nelson and he's from the Jewelry Tech Team. The man holds a red saw frame labeled 'Knew Concept'. The screen displays the text 'FROM THE JEWELRY TECH TEAM'. Another screen reads 'TECH REVIEW: KNEW CONCEPTS MK4 SAW FRAME'. The man starts demonstrating the saw's features. He holds a saw in one hand and a thin piece of copper in the other, gesturing at how to cut the piece of copper with the saw. The man picks up two saw frames of the same model but different sizes to compare. He puts on a magnifying eyewear, then takes it off and puts it on the table. The man shows how to cut copper with the saw. He wears magnifying eyewear while using the saw, which is clamped to a table. The video concludes with the Rio Grande logo.

Figure 15: Examples of minor inaccuracies in descriptions generated by VideoA11y Red underline indicates the inaccuracies content in the descriptions.