Mind the Gap: Dataset and Fine-grained Evaluation for Inline Audio Descriptions

Anonymous CVPR submission

Paper ID *****

Abstract

Audio descriptions (AD) are essential for making visual media accessible to blind and low-vision (BLV) users. While Multimodal Large Language Models (MLLMs) offer a scalable solution for generating audio descriptions on-demand, their performance on "inline" audio descriptions – which must fit within existing silences in a video – remains under-explored, particularly for diverse user-generated content (UGC). We present a comprehensive investigation into MLLM-generated inline audio descriptions. First, we introduce a professionally annotated dataset of 36 hours of videos, each averaging ∼5 mins. and total 7k+ descriptions, providing a high-quality benchmark for this task. Second, we propose an evaluation framework based on audio description expert guidelines that provides fine-grained actionable metrics for model improvement. Crucially, our framework evaluates the end-to-end task: it renders the generated scripts into audio tracks to assess their fit within the original video's natural silences, minimizing disruptive overlaps. Our analysis reveals that while frontier MLLMs accurately describe visual events, significant gaps persist in audio-visual timing, quality, and narrative flow. Finally, we validate the framework and conduct studies with BLV participants and sighted raters, finding that while users narrowly prefer human-authored audio descriptions, they still find MLLM-generated audio descriptions highly beneficial for comprehension. We provide guidance on the current capabilities of MLLMs for audio descriptions and release our evaluation suite to the community.

1. Introduction

Blind and low vision (BLV) audience members watch user-generated videos to learn new skills, engage in their hobbies, and stay up to date [16]. However, video creators often do not describe all of the important visual content required to understand the video such that they remain inaccessible to BLV audiences [16, 17]. Audio descriptions (AD), or narrations of important visual content that avoid overlap with the important audio content, can make videos accessible to BLV audience members [1, 29].

While high-quality audio descriptions are professionally crafted for film and TV [21], such descriptions are rarely present for user-generated videos [16]. Semi-automated [5, 17, 23, 35] and automated approaches [3, 8, 14, 15, 31, 33] to generate audio descriptions thus have the potential to make user-generated videos accessible to millions of blind and low vision audience members. Recent progress in multimodal large language models (MLLMs) has made it newly possible to generate audio description scripts with general purpose architectures. For example, Gemini-2.5 [7], GPT 5.2 [20], and Qwen-2-VL-Max [30] can take a video as input then directly provide an audio description script with timestamps and narration that can then be rendered into audio using speech to text. But, ***are MLLM-generated audio descriptions understandable and enjoyable for BLV audiences of user-generated videos?*** To answer this, it is essential to evaluate the full end-to-end process: generating the AD script, synthesizing the audio via Text-to-Speech (TTS), and analyzing how well the resulting audio track integrates into the natural silences of the original video.

Prior work evaluates such generated AD via comparison to gold standard AD via general-purpose and AD-specific reference-based metrics [9, 34], evaluation with intended audience members [14], or reference-free LLM-based evaluation [14]. However, unlike movies that have professionally-crafted gold standard AD [25], high-quality AD are rare for user-generated videos. Prior work generating AD for user-generated videos thus used synthesized AD [9] or novice-created AD [14] gold standards for comparison. However, it is unclear whether success on these reference-based comparisons will translate to high-quality viewing experiences, and the coarse metrics used for human and LLM-based reference-free evaluation (*e.g.*, quality, satisfaction) have started to saturate when compared to such novice ADs [14]. Further, it is challenging to know how to improve BLV user's viewing experience based on coarse reference-free metrics alone.

Our work makes three contributions: First, we *estab-*

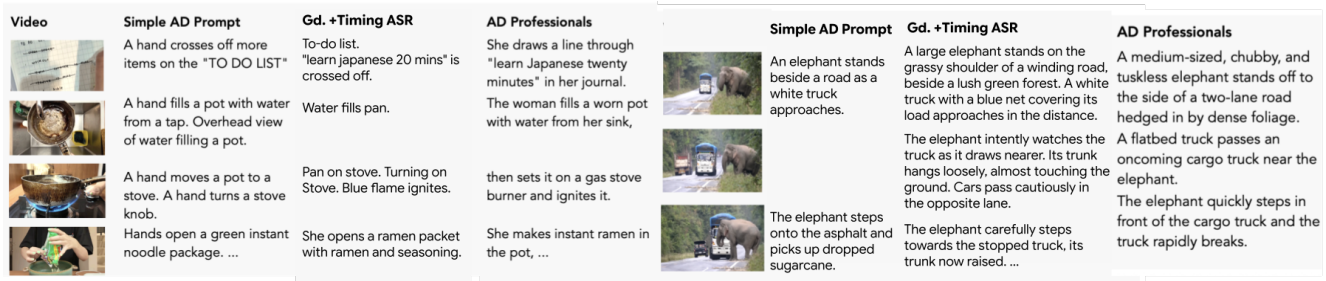


Figure 1. **Dataset Examples.** Model Generated AD (created with the Simple Prompt and Guideline+Timing ASR) and corresponding professionally annotated AD for a video segment. Here, we illustrate each video segment with manually selected video frames. Top example demonstrates narrative flow in the professionally annotated AD (describes woman's story rather than just the actions), and the example below shows a missed event in the Model Generated ADs (the truck rapidly breaks for the elephant).

lish a high-quality benchmark for AD of user-generated videos by collecting a dataset of professionally-crafted ADs for 400+ YouTube videos. To assure the dataset accurately reflects the video and the preferences of BLV audience members, we uniquely recruited pairs of BLV and sighted AD professionals who collaborated to craft each AD in our dataset. Second, we contribute an *evaluation framework that provides fine-grained and actionable metrics for model improvement* based on AD expert guidelines, including studying bias on LLM-as-a-judge, and showing generalization to existing datasets (YouDescribe). Finally, we *validate our dataset and framework* by conducting user studies with sighted raters and 8 BLV participants demonstrating that while users narrowly prefer human-authored AD, they find generated AD useful for comprehension. With our framework and user study findings, we provide guidance on opportunities to further improve MLLM generated ADs.

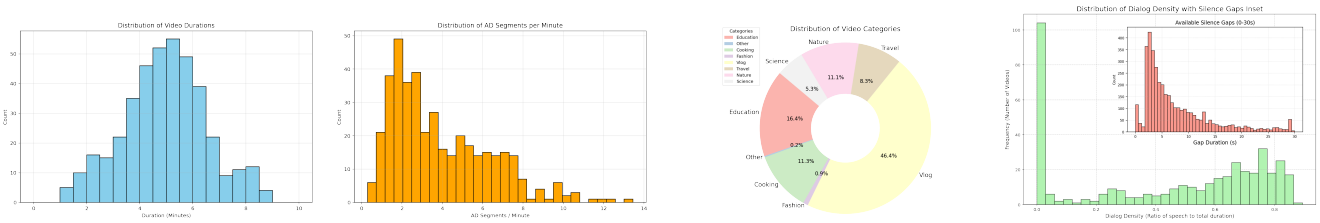
2. Related Work

We review prior AD datasets, approaches for evaluating AD, and AD generation techniques to motivate our dataset and evaluation framework for AD for user generated videos.

Audio Description Datasets. Prior work has proposed a number of datasets to support generating and evaluating AD, see Table 3 for a comparison. The majority of datasets use already existing professional ADs for movies [9, 25, 27, 34] or TV episodes [33]. These datasets of professionally-crafted AD thus can inform *what* should be described and *how* it should be described in films and TV episodes. However, the characteristics of movies and TV episodes (*e.g.*, intended for entertainment, many characters, highly curated shots, short gaps for AD) differ from the characteristics of user generated videos (*e.g.*, heterogeneous purposes, capture styles, editing styles, and narration density). Furthermore, understanding user-generated videos requires reasoning over long time horizons (averaging 5 minutes in our dataset), a challenge highlighted by recent long-video benchmarks such as Neptune [18]. As professional ADs rarely exist for user-generated videos, prior work collected novice-created descriptions of short 3-10s video clips of YouTube videos [14] and synthesized AD by adapting narrations from how-to videos [9] to serve as proxies for gold standard AD.

Evaluating Audio Descriptions. The primary way to evaluate AD is by automatically comparing generated AD to gold standard AD [6] using a range of general-purpose metrics from classic text similarity metrics such as BLEU [22] to embedding-based semantic similarity metrics such as BERTScore [36], or AD-specific metrics such as MNScore which checks for semantic similarity and character recognition [34], CRITIC [9] which assesses character recognition, and LLM-AD-Eval [9] which assesses semantic similarity. Our work creates a dataset that will enable such reference-based assessment for user-generated videos.

While such reference-based metrics can support assessing AD accuracy and similarity to the gold standard, such metrics do not necessarily correspond to the audience's understanding and enjoyment of AD. Recent work has highlighted the limitations of evaluating AD on short, trimmed clips using single ground-truth references, arguing for sequence-level narrative coherence [13] or Question Answering (QA) benchmarks to assess visual appreciation and narrative understanding [12]. Thus, prior work has invited external annotators or target audience members to evaluate the quality of AD — but, such evaluations of end-to-end AD are time-consuming and thus typically involve few video samples [2, 23]. Recent work invited blind audience members to evaluate quality of single AD segments (*e.g.*, descriptiveness, objectivity, accuracy, and clarity) for short ∼10s video clips [14] and found the ratings to be comparable to LLM ratings of the same metrics, and to be comparable between novice-created and automatically-created descriptions. The validity of using LLMs as judges for audio-related captions has also been strongly supported by recent studies demonstrating high correlation with human judgments [32]. However, prior work does not yet explore human-ratings or automated metrics to evaluate the perceived quality of end-to-end generated ADs of user-



(a) Video Durations (b) AD Segment Count (c) Video categories (d) Speech Density

Figure 2. **Dataset Statistics.** Our professionally annotated dataset consists of 438 videos (∼36 hours). (a) Most videos are centered around 5 mins; (c) Vlogs represent the largest category, along with travel, nature, education reflecting real-world user content; (b) distribution of human annotated AD segments per video; (d) the density of speech and the (inset) distribution of silence gaps highlights the tight temporal constraints of the inline AD task, where descriptions must fit within brief pauses in speech.

generated videos. Prior work also does not compare human ratings of professionally-crafted AD to generated AD to assess the remaining gap.

3. Mind the Gap Dataset

We aimed to establish a dataset of gold standard audio descriptions for user generated videos. We first selected a diverse sample of user-generated videos on YouTube as BLV audience members have expressed interest in watching YouTube videos [14, 16, 26]. We then hired audio description professionals who worked in teams of blind and sighted creators to provide descriptions for each video to ensure that our gold standard descriptions were accurate and reflected the preferences of the target audience. Finally, to promote comparison and future research, we add to our dataset end-to-end audio descriptions generated by the best prompt and model in our study, and for 100+ videos we add audio descriptions generated by several frontier multi-modal models that process videos. Figure 2 shows a summary of dataset statistics and Figure 1 shows some examples.

3.1. Video Selection

We selected **initial candidate videos** for annotation from YouTube by first identifying channels that had at least 1000 subscribers, then selecting videos from these channels that were around 3-7 minutes long to facilitate annotation and evaluation. Our initial set featured 557 videos. We then provided the videos to the professional AD team for the AD annotations and any further filtering.

The professional team's **video filtering process** began with a full review of each candidate video to assess its viability for audio description. A video was deemed non-viable if it contained nonstop dialogue or narration, leaving no temporal gaps for new descriptions, or if it was a vlog where existing subtitles occupied all available audio space. The process was iterative, with the team later revisiting some non-viable videos to add partial descriptions, concluding that some annotation was preferable to none. To ensure dataset diversity, a redundancy protocol was established; after annotating 3–5 videos of a similar theme (e.g., cooking tutorial having a common style), subsequent similar videos were skipped. Ultimately, 438 videos (78.6%), accounting for over 36 hours of content, were successfully annotated. 92 videos (16.5%) were excluded as non-viable, and an additional 27 (4.6%) were skipped due to redundancy. Figure 2 shows the properties of the user generated videos in our dataset (*e.g.*, around 5 minutes, from a diversity of categories, with different levels of dialogue density).

3.2. Professionally Annotated AD

The professional descriptions were curated through a meticulous, collaborative **annotation process** executed by two (two-person) teams, each comprising one blind and one sighted writer. The primary annotation was performed by the sighted writer, who began by watching each video in its entirety to grasp the overall content and pacing. Following this, they rewatched the video in segments, pausing to draft concise, time-aligned descriptions of visual information not conveyed by existing dialogue or sound. This included transcribing on-screen text when present. A critical step involved importing the drafted script into an editing environment to test and refine the timing of the descriptions against the video's original audio track. This ensured the added descriptions fit naturally within available pauses, avoiding overlap with dialogue and maintaining a smooth, unobtrusive viewing experience. Finally, the completed script was submitted to the blind writing partner for a comprehensive quality control review, leveraging their expertise to validate the clarity and effectiveness of the descriptions. Figure 2 demonstrates the number of AD segments per video with a median around 2 segments per minute, for an average of around 18 AD segments per video, and 7325 AD segments across the entire dataset.

3.3. Model Generated AD

We also obtained audio descriptions from closed source multimodal models that process videos. We used 4 different prompting strategies, ranging from simple to more comprehensive and reasoning oriented to capture a range of descriptions. The **Simple Prompt** (in App. Fig. 14) very briefly instructs the model to generate inline audio descrip-

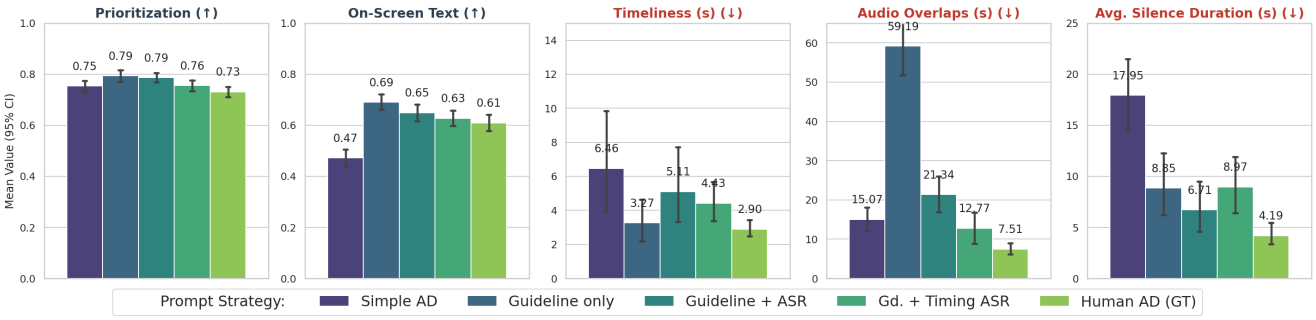


Figure 3. **Prompt Strategy** comparisons for AD generation. **Content:** The first two panels (Higher is Better) demonstrate that automated AD effectively captures visual prioritization. Further, guideline-based prompts significantly improve narration of relevant information such as on-screen text. **Temporal precision** The latter three panels (Lower is Better) highlight the trade-offs in temporal alignment. While "Simple AD" and "Guideline only" prompts result in excessive overlaps and poorly managed silence, the Gd. Timing+ASR strategy—which integrates guidelines with transcript reasoning—successfully balances content richness with strict temporal constraints.

tions for a given video and specifies the output format. A second prompt (**Guideline AD**, App. Fig. 15) provides a more detailed description of what is expected in an audio description, including several points from the guidelines for AD. A third more informative prompt (**Guideline + ASR**, App. Fig. 16) builds on the Guideline AD prompt by also including the transcript and timestamps for the speech in the original videos. The final more comprehensive prompt (**Gd. + Timing ASR**, App. Fig. 17), in the flavor of chain-of-thought, further adds information about available duration and time to speak out content to allow the model to reason about the length and placement of the descriptions.

4. Development of Metrics

Expert-Informed Guidelines. To create an evaluation framework for audio descriptions (AD) that provides interpretable and actionable results for model improvement, we reviewed a set of expert audio description guidelines from five sources selected for comprehensiveness and diversity: the American Council of the Blind's guidelines [1], DCMP's Description Key [4], ADLab's guide for describing film [24], W3C's Web Content Accessibility Guidelines [28], and YouDescribe's tips for description [26], following a process similar to prior work [14, 23]. We also reviewed literature on BLV audience members' perceptions of AD to surface deviations between expert best practices and specific audience preferences [10, 11, 16, 19].

In collaboration with researchers in Human-Computer Interaction (HCI), Accessibility, and Computer Vision (CV), we synthesized these findings into general-purpose best practices (guidelines G1-G10, see Appx. Sec. 6 for a full list of guidelines). To ensure the framework addresses real-world needs, we prioritized these guidelines based on a pilot study where BLV audience members reviewed descriptions created with a basic prompt (e.g., "Create an audio description"). During this pilot, users surfaced the issues that most negatively impacted their experience (*e.g.*, AD overlapping audio, empty silences, misaligned descriptions), and these insights informed our selection of guidelines for the final evaluation. Consequently, we developed a suite of quantifiable metrics to evaluate AD across three primary dimensions: Content Accuracy, Temporal Precision, and Qualitative Alignment.

4.1. Metrics for Evaluation

We propose and employ a hybrid suite of metrics for the evaluation of generated audio descriptions. To ground our metrics, we utilize the synthesized expert guidelines (G1-G10) encompassing content (G1-G4), style (G5-G6), and audio quality (G7-G10). For instance, G1 focuses on Content Coverage, G2 on Timeliness, and G7 on minimizing Audio Overlaps (see Appx. Sec. 6 for the full list). From these, we derived specific quantifiable metrics. Some metrics are measured precisely and programmatically and others use an LLM-based prompt to extract more detailed scores which we then parse and aggregate for each video and across videos.

Content Accuracy Metrics These metrics assess the "what" of the description, ensuring the most relevant visual information is captured.

- **Prioritization (G1.1):** Measures how effectively the generated AD covers critical visual content. We utilize an MLLM to compare the generated AD against dense time-based visual information, calculating a coverage score (recall) based on the presence of key visual events. App. Fig. 13 shows how we get dense visual information and App. Fig. 18 shows the prompt used to elicit prioritization of events in the AD.
- **Redundancy (G1.2):** AD does not describe information that is already clear to BLV viewers from the audio alone (*e.g.*, redundant with speech or sounds). To measure this, we provided an automated speech recognition (ASR) transcript programmatically interleaved with the generated AD based on the timestamp information to the LLM with a prompt (App. Fig. 20) to assess how redundant each AD line is with the surrounding transcript on a

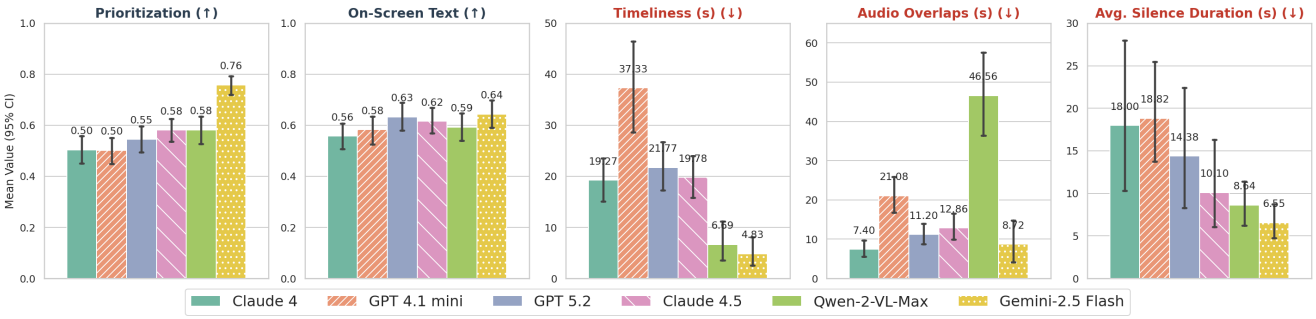


Figure 4. **Frontier model comparison. Content Accuracy** (First two plots): Higher scores indicate better prioritization and text recognition. **Temporal Precision** (Last three plots): Lower values indicate better performance (less delay, overlap, and silence). Gemini-2.5 Flash and Qwen-2-VL-Max demonstrate superior timing with the lower gaps of silence, and better prioritization of content.

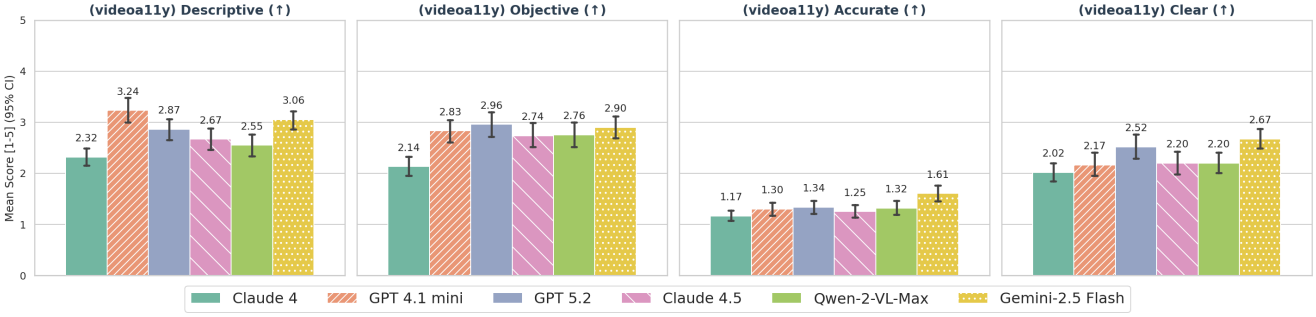


Figure 5. **LLM (VideoA11y) rating w.r.t Human Audio Descriptions.** Mean and 95% CI on a 1–5 scale of LLM-generated audio descriptions in comparison with human expert audio descriptions using the VideoA11y prompt [14] to measure Descriptiveness, Objectiveness, Accuracy, and Clarity (Higher is better).

scale from 0 (not redundant) to 1 (fully redundant). We then calculated the mean of the values to for an overall redundancy score.

- **On-Screen Text (G1.3):** Quantifies the narration of text appearing within the video frame. We first extract all on-screen text snippets and their timestamps programmatically from the dense visual information. Then this metric is calculated as the mean ratio of words from these snippets successfully incorporated into the AD.

Temporal Precision Metrics Inline AD must fit within the natural pauses of the original video audio. We define three metrics to measure this "fitness". For two of them, we do so by rendering the AD using an off-the-shelf Text-to-Speech (TTS) module at a fixed speaking rate of 1.2 and compute the duration of each AD.

- **Timeliness (G2):** Measures the absolute temporal offset (in seconds) between a visual event's appearance and its corresponding description. Lower values indicate better synchronization with the visual flow. To measure this we provide the LLM the visual information along with the generated AD to assess how far the time of the generated AD occurs from the occurrence of the visual information in the video (prompt in App. Fig. 21). We calculate the mean and max absolute offset to achieve a timeliness score.
- **Audio Overlaps (G7):** Calculates the total duration (seconds) where the generated AD (rendered using TTS) overlaps with the original video speech (determined via ASR transcript timestamp). Minimizing overlaps is critical for maintaining the intelligibility of the content.
- **Avg. Silence Duration:** Measures the remaining "dead air" in the video after AD is inserted. With the TTS rendered AD, we compute placement in the timeline based on the transcript timestamps and measure gaps of silence. While some silence is necessary, high values suggest under-description or poor utilization of available gaps.

Qualitative Alignment (VideoA11y) Following expert-informed best practices and adopting the VideoA11y framework proposed by Li et al. [14], we evaluate the stylistic quality of the audio descriptions on a 1-5 Likert scale using an LLM-based judge (Gemini, Qwen, or GPT) relative to human ground truth descriptions. These metrics measure alignment with specific professional standards:

- **Descriptive:** The extent to which the AD provides vivid, specific visual detail.
- **Objective:** Whether the AD describes visual facts without subjective interpretation or censorship.
- **Accurate:** The factual correctness of the AD.
- **Clear:** Logical flow and linguistic clarity of narration.

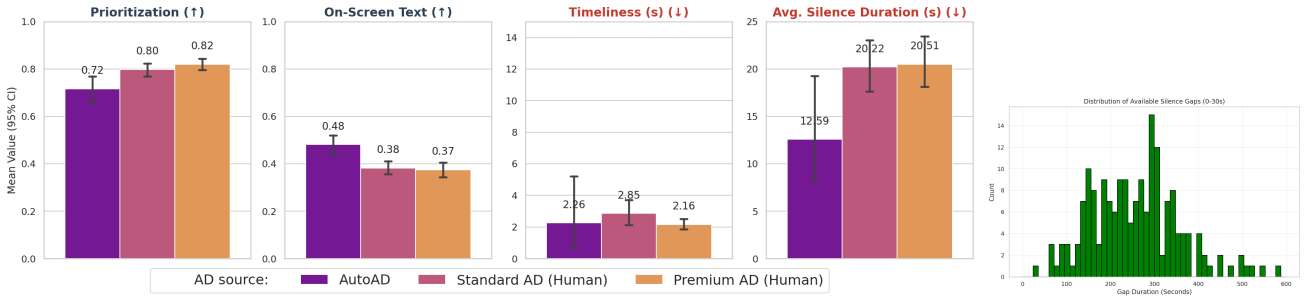


Figure 6. **YouDescribe generalization.** (a): Mean values (with 95% CI) for prioritization, on-screen text coverage, timeliness, and silence duration across 2 available human AD sources and our automated AD on the YouDescribe dataset. (b)Dist. of available silence gaps (right) shows significantly longer gaps compared to our dataset due to the source videos being largely silent or lacking pre-existing dialogue.

5. Experiments and Results

We evaluate our proposed framework through three lenses: (1) automated metric-based evaluation of prompting strategies and frontier MLLMs; (2) a study on LLM-as-a-judge reliability and generalization to existing datasets; and (3) user studies with sighted raters and blind and low-vision (BLV) participants.

5.1. Metric-based Evaluations

Prompting Strategies. We first compare four prompting strategies using our automated metrics (Fig. 4, App. Tab. 1). We find a significant trade-off between content richness and temporal precision. While basic prompts achieve high Prioritization scores ($0.75 - 0.79$), they fail severely on temporal alignment. Specifically, the *Simple AD* and *Guideline only* prompts result in excessive silent gaps or overlaps ($> 50s$). In contrast, the *Gd. + Timing ASR* strategy, which incorporates ASR transcripts and temporal reasoning, achieves the best balance, reducing overlaps to $12.77s$ and timeliness offsets to $4.43s$ while maintaining high prioritization (0.76).

Frontier Model Comparison. Using the *Gd. + Timing ASR* prompt, we benchmark several MLLMs (Fig. 4, App. Tab. 2). Gemini-2.5 Flash and Qwen-2-VL-Max demonstrate superior performance. Gemini-2.5 Flash achieves substantially higher prioritization (0.76) and effective on-screen text recognition (0.64), and along with Qwen-2-VL-Max show good temporal precision and low gaps of silence.

Qualitative Alignment (VideoA11y). We further evaluate models on stylistic metrics using the VideoA11y suite (Fig. 4). While models perform relatively well on Objectivity ($2.14-2.96$) and Descriptiveness ($2.32-3.24$), there remains a significant gap compared to Human Ground Truth, particularly in Accuracy and Clarity, where most models score below 3.0 on a 5-point scale. This used Gemini 2.5 Flash as a judge.

5.2. LLM as Judge and Generalization

Inter-judge Reliability. To validate the use of MLLMs as evaluators, we measure the correlation between Gemini, Qwen, and GPT-4o judges across five metrics. As shown in Fig. 7, GPT-4.1 Mini and Qwen-VL-Max show high alignment. Bias analysis (Fig. 7b) reveals that while some models are "lenient" (e.g., GPT-4.1 Mini), others are "harsher". We hypothesize that models with stronger reasoning capabilities may better identify subtle hallucinations or lack of detail that smaller models miss, leading to lower scores. However, the systematic deviation remains low, supporting the use of MLLMs for scalable AD evaluation.

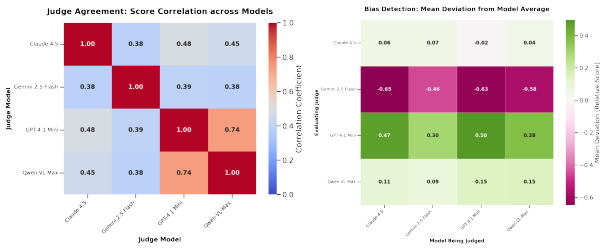


Figure 7. Judge Alignment and Bias. (a) Inter-judge agreement on prioritization and video accessibility metrics. Pearson correlation of average scores across 100+ videos. Each score is an aggregate of five key accessibility metrics (e.g., precision of visual information, and videoa11y based descriptiveness, clarity etc.) Here GPT-4.1 Mini and Qwen-VL-Max show high alignment. (b) Deviation from model-specific mean scores. Shows systematic bias of each judge by calculating the mean deviation from the global average score per video. Positive values (green) indicate "lenient" judging, while negative values (pink) indicate "harsh" judging relative to the consensus.

Generalization to YouDescribe. We apply our framework to the YouDescribe dataset to assess its robustness across human-generated audio descriptions of varying quality (Fig. 6). Our metrics successfully distinguish between "Premium" and "Standard" human descriptions, with Premium descriptions showing better timeliness ($2.18s$ vs. $2.85s$) and higher prioritization (0.82 vs. 0.80). These results validate that our proposed metrics align with human quality judgments even on external datasets, suggesting their utility as a universal benchmark for audio description quality.

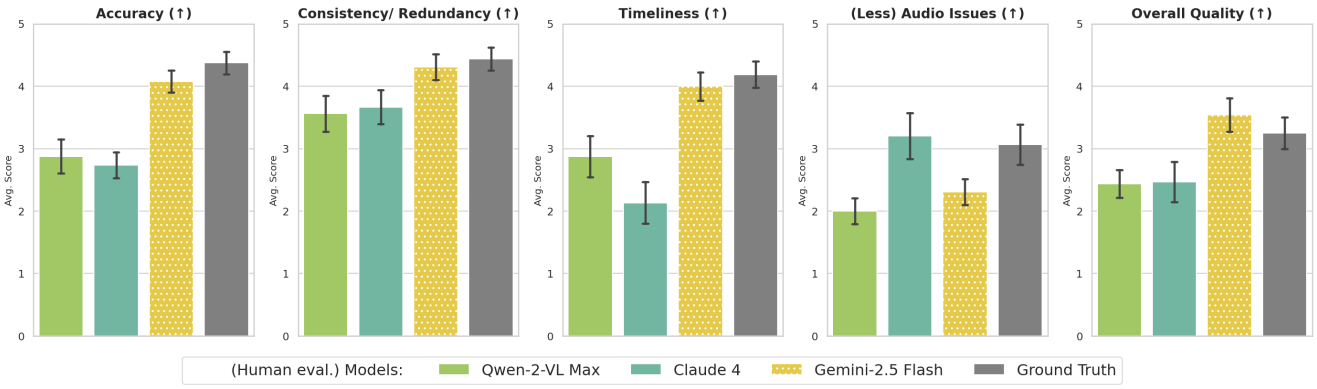


Figure 8. **Sighted rater model comparisons**. Mean Likert scores across five (of 11) key dimensions on videos evaluated by sighted raters (↑ indicates higher is better). Gemini consistently performs closest to the Ground Truth across Accuracy, Consistency, and Timeliness.

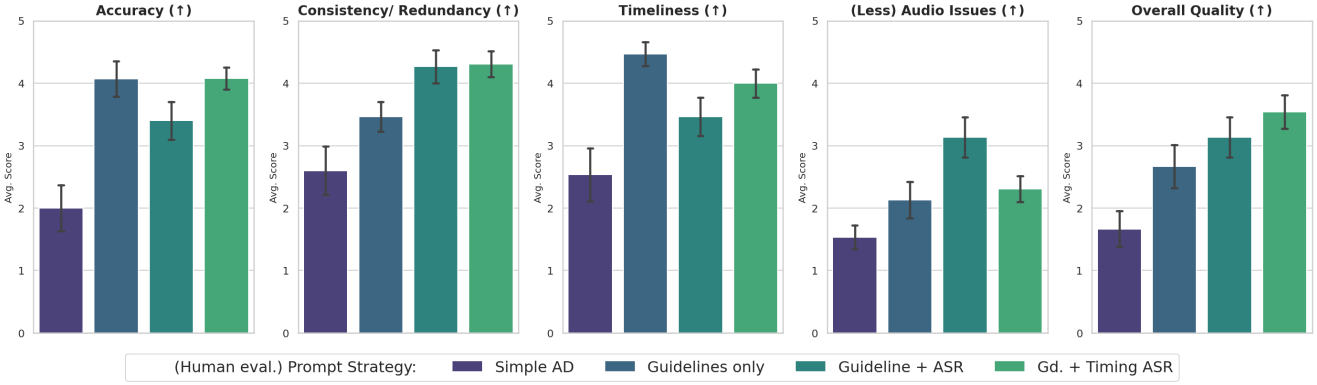


Figure 9. **Sighted rater prompt comparisons**. Mean Likert scores across five (of 11) key dimensions (↑ indicates higher is better). Sighted raters evaluated videos played with AD generated from different prompts.

5.3. Sighted and BLV Participants Studies

We conducted human studies with both sighted raters and BLV participants using a custom UI where the raters could watch the video and listen to the rendered audio descriptions, and provide ratings on a likert scale (1-5).

Sighted Rater Study. A pool of 8 sighted raters evaluated 42 videos across 11 dimensions including 'Coverage', 'Accuracy', 'Relevance', 'Clarity', 'Timeliness', and 'Audio Issues' (full list of questions and rating options are in App. Sec. 6). Results in Fig. 8 and Fig. 9 confirm that Gemini-2.5 Flash with the *Timing ASR* prompt performs closest to human ground truth in Accuracy and Timeliness. Inter-annotator agreement (Krippendorff's α) across all metrics was 0.654 and was highest for Timeliness (0.721) and Audio Issues (0.700). Scores in Appx. Tab. 4.

BLV Participant Study. We recruited 7 BLV participants who use YouTube. Each participant listened to AD for four videos randomly sampled from our dataset. For each video, participants listened to both automatically generated AD (Gemini-2.5 Flash with *Gd. + Timing ASR*) and the human-generated AD created by blind and sighted AD professionals. Participants were unaware of the AD source and we randomized and counterbalanced the order of the AD sources. Participants provided ratings for each AD (Fig. 10), their preference between the two AD (Fig. 11), and qualitative feedback about each AD.

Participants rated both versions consistently high across all dimensions (Fig. 10). While they marginally preferred human-authored AD (53.6%) (Fig. 11), they found MLLM-generated AD highly beneficial for video comprehension. Specifically, generated AD scored near-parity with human experts in Quality, Clarity, and Objectivity. Although human-crafted AD scored slightly higher across all ratings, significance testing via the Wilcoxon Signed Rank Test (due to use of ordinal data) revealed that BLV participants rated professionally crafted AD higher for only their overall Satisfaction ($p < 0.05$) with the descriptions and their perceived Accuracy ($p < 0.05$). Qualitative feedback indicated that BLV users preferred professional AD for their superior "narrative flow" and concrete details that complemented rather than repeated the audio. In contrast with prior work that found that generated AD was rated consistently higher than human novice-crafted AD for short video clips [14],

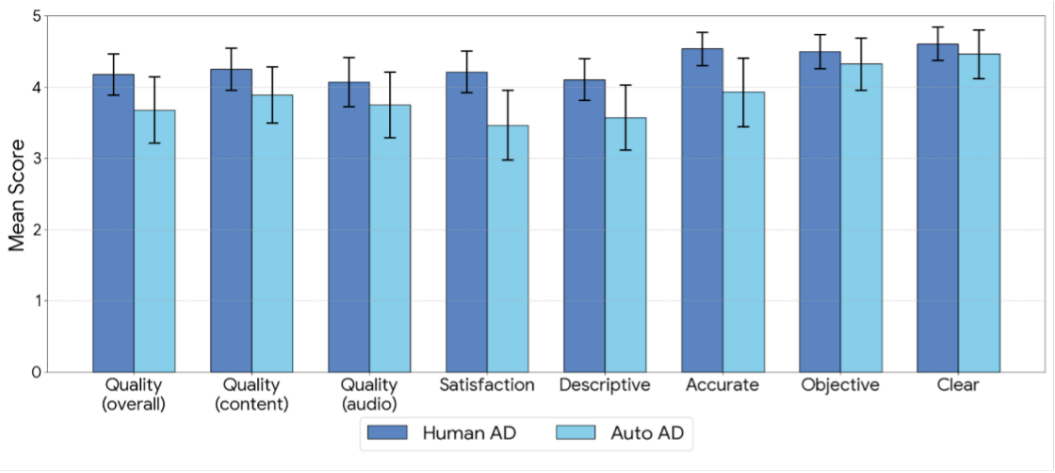


Figure 10. **BLV study**. Mean Likert scores across 8 key dimensions (error bars depict 95% confidence intervals). BLV participants evaluated videos played with human written ADs and AutoAD (Gemini-2.5 Flash with Gd. + Timing ASR). They rated both AD versions consistently high across all metrics and the human generated AD slightly consistently better in all dimensions.

our results indicate that even though our generated AD received similar ratings for longer video segments, a gap still remains between generated AD and our new dataset of gold standard professionally-crafted AD.

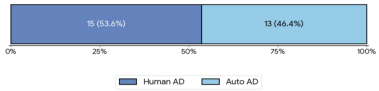


Figure 11. **BLV Study.** In a blind study, BLV participants listening to the same video with human generated AD and AutoAD (Gemini-2.5 Flash with Gd. + Timing ASR) when asked to pick which AD version they liked better, marginally preferred the video version with human generated AD.

6. Conclusion and Broader Impact

Our investigation establishes that while frontier multimodal large language models (MLLMs) have reached a "near-parity" milestone in basic visual descriptiveness, the "gap" in professional-grade inline audio description (AD) is defined by temporal precision and narrative artistry. Through the *Mind the Gap* dataset—the first of its kind to be co-created by collaborative teams of blind and sighted professionals—we provide the community with over 36 hours of high-quality ground truth for the challenging domain of user-generated video.

Our multi-dimensional evaluation reveals 3 key findings:

- **Prompting for Precision:** Incorporating ASR transcripts and temporal reasoning (via the *Gd. + Timing ASR* strategy) is essential for balancing content richness with strict temporal constraints, reducing audio overlaps.
- **Model Performance:** Frontier models like Gemini-2.5 Flash and Qwen-2-VL-Max demonstrate superior capabilities in content prioritization and timing, though they still struggle with the factual accuracy and linguistic clarity found in human-authored scripts.
- **User Utility:** Our study with blind and low-vision (BLV) participants confirms that while human-crafted AD remains the gold standard for narrative flow, MLLM-generated AD is already highly beneficial for comprehension, scoring consistently high across quality and clarity dimensions.

By releasing our evaluation suite and dataset, we aim to catalyze the development of AD systems that do not just describe pixels, but weave them into the cohesive, enjoyable narratives that all audiences deserve.

Limitations While our work introduces a comprehensive framework, several limitations remain. First, our automated metrics currently prioritize a subset of prioritized guidelines (G1.1-1.3, G2, G7, G8) based on pilot studies, meaning aspects like active voice (G5) and prosody (G10) are not yet programmatically evaluated. Second, our user studies, while insightful, involved a relatively small sample size of individuals, which may not capture the full diversity of preferences within the BLV community. Finally, the current framework relies on off-the-shelf text-to-speech (TTS) modules for temporal assessment, which may not fully reflect the nuances of professional narration styles.

Broader Impact The *Mind the Gap* dataset and evaluation framework have the potential to significantly democratize video accessibility. By providing a scalable way to generate and audit inline AD, this work can help make millions of hours of user-generated content—from educational tutorials to personal vlogs—accessible to the BLV community. Beyond direct accessibility, the metrics proposed here could serve as real-time "accessibility nutrition labels," informing all users about the quality of descriptions before they engage with content. Ultimately, this research moves us closer to a future where visual information is universally perceivable through high-quality, automated narration.

References

[1] ACB. American council of the blind, audio description project, guidelines for audio describers. https://www.acb.org/adp/guidelines.html, 2020. 1, 4

[2] Carmen J Branje and Deborah I Fels. Livedescribe: can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness*, 106(3):154–165, 2012. 2

[3] Peng Chu, Jiang Wang, and Andre Abrantes. Llm-ad: Large language model based audio description system. *arXiv preprint arXiv:2405.00983*, 2024. 1

[4] DCMP. Description key. https://dcmp.org/learn/captioningkey/624, 2020. 4

[5] Langis Gagnon, Samuel Foucher, Maguelonne Heritier, Marc Lalonde, David Byrns, Claude Chapdelaine, James Turner, Suzanne Mathieu, Denis Laurendeau, Nath Tan Nguyen, and et al. Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Universal Access in the Information Society*, 8(3):199–218, 2009. 1

[6] Yingqiang Gao, Lukas Fischer, Alexa Lintner, and Sarah Ebling. Audio description generation in the era of llms and vlms: A review of transferable generative ai technologies. *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 471–490, 2025. 2

[7] Google DeepMind. Gemini 2.5: Enhanced reasoning and multimodal capabilities. Technical report, Google, 2025. Gemini 2.5 Technical Report. 1

[8] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940, 2023. 1, 12

[9] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad iii: The prequel-back to the pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18164–18174, 2024. 1, 2, 11, 12

[10] Lucy Jiang, Mahika Phutane, and Shiri Azenkot. Beyond audio description: Exploring 360° video accessibility with blind and low vision users through collaborative creation. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*, 2023. 4

[11] Lucy Jiang, Crescentia Jung, Mahika Phutane, Abigale Stangl, and Shiri Azenkot. "it's kind of context dependent": Understanding blind and low vision people's video accessibility preferences across viewing scenarios. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2024. 4, 11

[12] Divy Kala, Eshika Khandelwal, and Makarand Tapaswi. What you see is what you ask: Evaluating audio descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. 2

[13] Eshika Khandelwal, Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Andrew Zisserman, Gül Varol, and Makarand Tapaswi. More than a moment: Towards coherent sequences of audio descriptions. *arXiv preprint arXiv:2510.25440*, 2025. 2

[14] Chaoyu Li, Sid Padmanabhuni, Maryam S Cheema, Hasti Seifi, and Pooyan Fazli. Videoa11y: Method and dataset for accessible video description. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–29, 2025. 1, 2, 3, 4, 5, 7, 12

[15] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*, 2023. 1

[16] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. What makes videos accessible to blind and visually impaired people? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–4, New York, NY, USA, 2021. ACM. 1, 3, 4, 11

[17] Xingyu Liu, Ruolin Wang, Dingzeyu Li, Xiang'Anthony' Chen, and Amy Pavel. Crossa11y: Identifying video accessibility issues via cross-modal grounding. In *UIST 2022*, 2022. 1

[18] Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, Mikhail Sirotenko, Yukun Zhu, and Tobias Weyand. Neptune: The long orbit to benchmarking long video understanding. In *arXiv preprint arXiv:2412.09582*, 2024. 2

[19] Rosiana Natalie, Ruei-Che Chang, Smitha Sheshadri, Anhong Guo, and Kotaro Hara. Audio description customization. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–19, 2024. 4, 11

[20] OpenAI. Gpt-5.2 system card. Technical report, OpenAI, 2025. Technical Report. 1

[21] Jaclyn Packer, Katie Vizenor, and Joshua A Miele. An overview of video description: history, benefits, and guidelines. *Journal of Visual Impairment & Blindness*, 109(2): 83–93, 2015. 1

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 2

[23] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (Virtual Event, USA) (UIST '20)*, pages 747–759, New York, NY, USA, 2020. Association for Computing Machinery. 1, 2, 4

[24] A. Remael, N. Reviers, and G. Vercauteren. Pictures painted in words: Adlab audio description guidelines. https://dcmp.org/learn/captioningkey/624. 4

[25] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 1, 2, 12

[26] SKERI. Youdescribe. the smith-kettlewell eye research institute. https://youdescribe.org/, 2022. 3, 4, 12

[27] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022. 2, 12

[28] W3C. G8: Providing a movie with extended audio descriptions. https://www.w3.org/TR/WCAG20-TECHS/G8.html. 4

[29] W3C. Audio description (prerecorded): Understanding sc 1.2.5. https://www.w3.org/TR/UNDERSTANDING-WCAG20/media-equivaudio-desc-only.html, 2022. 1

[30] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Junyang Zhou, Jingren Zhou, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[31] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 2021. Association for Computing Machinery. 1

[32] Tsung-Han Wu, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Claira: Leveraging large language models to judge audio captions. In *arXiv preprint arXiv:2409.12962*, 2024. 2

[33] Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad-zero: A training-free framework for zero-shot audio description. In *Proceedings of the Asian Conference on Computer Vision*, pages 2265–2281, 2024. 1, 2

[34] Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. Movie101: A new movie understanding benchmark. *arXiv preprint arXiv:2305.12140*, 2023. 1, 2, 12

[35] Beste F 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. Human-in-the-loop machine learning to increase video accessibility for visually impaired and blind users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 47–60, New York, NY, USA, 2020. Association for Computing Machinery. 1

[36] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 2

Appendix

Synthesized Suite of metrics from Expert Guidelines

This section presents the full suite of evaluation metrics we synthesized based on expert guidelines.

AD Script Content Quality:

- **G1 - Content Coverage:** AD describes important visual content necessary to understand the video.
  - **G1.1 - Prioritization:** AD prioritizes visual information necessary to understand the video.
  - **G1.2 - Redundancy:** AD does not describe information that is already clear to BLV viewers from the audio alone (*e.g.*, redundant with speech or sounds).
  - **G1.3 - On-Screen Text:** AD narrates on-screen text.
  - **G1.4 - Visual Style:** AD describes style, color, textures and temporal transitions only as useful to understand the video (audience members differ in their preferences for descriptions around visual styles [11, 16, 19]).
- **G2 - Content Timeliness:** AD describes visual content close to the time it appears on screen.
- **G3 - Content Accuracy:** AD contains only accurate information (similar metrics already addressed in prior work [9]).
- **G4 - Content Objectivity:** AD describes content as it appears on screen and does not include censorship or interpretation (audience members differ in their preferences for objective vs. subjective descriptions [11, 16]).

AD Script Style:

- **G5 - Active Voice:** AD script is written in third person with active voice.
- **G6 - Appropriateness:** Language of the AD matches the language of the video such that it is appropriate for the audience.
  - **G6.1 - Consistency:** AD uses language that is consistent within the AD and with the speech in the video.
  - **G6.2 - Terminology and Tone:** AD uses vocabulary and tone that are similar to the vocabulary and tone in the video (*e.g.*, does not introduce jargon, uses similar level of formality).

AD Audio Quality:

- **G7 - Overlaps**: AD avoids overlapping speech.
- **G8 - Audio Coverage**: AD covers the majority of the video, but does not need to fill every silence.
- **G9 - Pronunciation**: AD pronounces words correctly.
- **G10 - Appropriate Speech Style:** Speed and prosody are appropriate for the content of the video (*e.g.*, somber for a melancholy scene).

For the analysis of the main experiments in this work, we use a subset of our guidelines that we prioritized for evaluation based on a pilot study with BLV audience members reviewing descriptions created with a basic prompt (*e.g.*, "Create an audio description"). In the pilot study BLV users surfaced issues that most negatively impacted their experience, and these issues informed our choice of guidelines for initial metrics (**G1.1-1.3**, **G2**, **G7**, **G8**).

Full set of experimental results on proposed metrics.

Full experimental results on all metrics

Metric	Simple AD	Guideline only	Guideline + ASR	Gd. + Timing ASR	Human AD (GT)
Prioritization (↑)	0.75 ±0.02	0.79 ±0.02	0.79 ±0.02	0.76 ±0.02	0.73 ±0.02
No Redundancy (↑)	0.77 ±0.03	0.72 ±0.02	0.71 ±0.02	0.69 ±0.03	0.72 ±0.02
On-Screen Text (↑)	0.47 ±0.03	0.69 ±0.03	0.65 ±0.03	0.63 ±0.03	0.61 ±0.03
Content Timeliness (s) (↓)	6.46 ±3.03	3.27 ±1.29	5.11 ±2.07	4.43 ±1.17	2.90 ±0.48
Overlaps w. TTS (s) (↓)	15.07 ±3.01	59.19 ±7.28	21.34 ±4.64	12.77 ±3.88	7.51 ±1.42
Audio Coverage (ratio) (↑)	0.41 ±0.06	1.23 ±0.11	0.83 ±0.08	0.78 ±0.05	0.61 ±0.03
Average Silence Duration (s) (↓)	17.95 ±3.46	8.85 ±3.13	6.71 ±2.47	8.97 ±2.80	4.19 ±1.13
VideoA11y Descriptive (↑)	3.46 ±0.13	3.28 ±0.14	2.81 ±0.11	3.03 ±0.10	4.72 ±0.07
VideoA11y Objective (↑)	3.55 ±0.13	3.59 ±0.13	3.27 ±0.12	2.98 ±0.13	4.93 ±0.03
VideoA11y Accurate (↑)	1.82 ±0.11	2.18 ±0.12	1.74 ±0.10	1.64 ±0.09	4.98 ±0.02
VideoA11y Clear (↑)	2.59 ±0.13	2.85 ±0.14	2.57 ±0.12	2.68 ±0.11	4.96 ±0.03

Table 1. **Prompt Comparisons**. Comparing performance on the proposed metrics on AD generated by different prompts. Showing the mean and the 95% confidence interval, ↑ (higher is better); ↓ (lower is better). Cells are shaded by rank: First and Second .

Metric	Gemini-2.5 Flash	Qwen-2-VL-Max	GPT 4.1 mini	Claude 4	Claude 4.5	GPT 5.2
Prioritization (↑)	0.76 ±0.04	0.58 ±0.05	0.50 ±0.05	0.50 ±0.05	0.58 ±0.05	0.55 ±0.05
No Redundancy (↑)	0.70 ±0.05	0.71 ±0.04	0.70 ±0.05	0.70 ±0.05	0.70 ±0.04	0.73 ±0.04
On-Screen Text (↑)	0.64 ±0.05	0.59 ±0.05	0.58 ±0.05	0.56 ±0.05	0.62 ±0.05	0.63 ±0.05
Content Timeliness (s) (↓)	4.83 ±2.96	6.69 ±3.76	63.06 ±17.92	19.27 ±4.26	19.78 ±4.19	21.77 ±4.64
Overlaps w. TTS (s) (↓)	8.72 ±5.41	46.56 ±10.90	21.08 ±4.61	7.40 ±2.06	12.86 ±3.22	11.20 ±2.67
Audio Coverage (ratio) (↑)	0.74 ±0.07	0.95 ±0.15	1.48 ±0.45	0.63 ±0.06	0.71 ±0.07	0.67 ±0.06
Average Silence Duration (s) (↓)	6.55 ±2.08	8.64 ±2.68	18.82 ±5.69	18.00 ±8.46	10.10 ±5.64	14.38 ±7.04
VideoA11y Descriptive (↑)	3.06 ±0.19	2.55 ±0.21	3.24 ±0.24	2.32 ±0.17	2.67 ±0.21	2.87 ±0.21
VideoA11y Objective (↑)	2.90 ±0.22	2.76 ±0.25	2.83 ±0.22	2.14 ±0.20	2.74 ±0.23	2.96 ±0.25
VideoA11y Accurate (↑)	1.61 ±0.15	1.32 ±0.13	1.30 ±0.13	1.17 ±0.10	1.25 ±0.13	1.34 ±0.14
VideoA11y Clear (↑)	2.67 ±0.20	2.20 ±0.20	2.17 ±0.23	2.02 ±0.18	2.20 ±0.22	2.52 ±0.24

Table 2. **Model Comparisons**. Comparing different frontier models on the performance of proposed metrics. Showing the mean and the 95% confidence interval, ↑ (higher is better); ↓ (lower is better). Cells are shaded by rank: First and Second .

Dataset	Type of Video	Total Hours	Avg Video Length	Source of AD	# of AD Segments	Fully Silent vs Dialogs	Eval. segment vs full
Videoa11y [14]	User-generated	N/A (40k videos)	∼10s	MLLMs (Synthetic)	40,000	Silent	Segment
YouDescribe (YuWA) [26]	YouTube	∼310 hours	∼5 mins	Volunteers / Novices	76,000	Both	N/A
M-VAD [25]	Movies	∼56.5 hours	6.2s	Professionals (DVS)	∼49,000	Mixed	Segment (gaps)
MAD [27]	Movies	1,207.3 hours	4.1s	Professionals (DVS)	> 384,000	Mixed	Segment (gaps)
Movie101v2 [34]	Movies (ZH/EN)	353 hours	11s - 20s	Professionals (DVS)	71,000	Silent	Segment (gaps)
CMD-AD [8]	Movie Clips	N/A (1,432 movies)	∼2 mins	Professionals (DVS)	∼101,000	Mixed	Segment (gaps)
HowTo-AD [9]	Instructional	∼180k videos	Variable	Synthetic	3.4 Million	Heavy dialog	N/A
Mind The Gap (Ours)	**User-generated**	**36 hrs**	**∼5 mins**	**Professional+Blind team**	**7325**	**Mixed**	**Full**

Table 3. **Comparison of popular Audio Description Datasets.** Comparing the kinds of videos, number of hours, source of annotation (by professionals or novices), presence of dialogs, and whether the evaluation was on silent segments alone or if it was end-to-end. Our dataset was sourced from professional AD teams consisting of one sighted and one blind AD creator working in pairs. Our evaluation is also on the full end-to-end AD generation for the entire video clip as opposed to each silent segment gap.

Inter Annotator Agreement

Full intera-annotator agreement on all of the metrics is shown in Table 4

Metric	Krippendorff's α	Avg Std Dev
Timeliness	0.721	0.581
Audio Issues	0.700	0.545
Objective	0.665	0.437
Gaps of Silence	0.647	0.637
Overall Quality	0.581	0.681
Descriptive	0.579	0.699
Relevance	0.563	0.641
Accuracy	0.547	0.706
Clarity	0.544	0.709
Coverage	0.539	0.700
Consistency/ Redundancy	0.537	0.696

Table 4. Inter-Annotator Agreement (Krippendorff's α) and Average Standard Deviation per Metric. Overall (all metrics): 0.654

Full distribution of scores from the BLV user study.

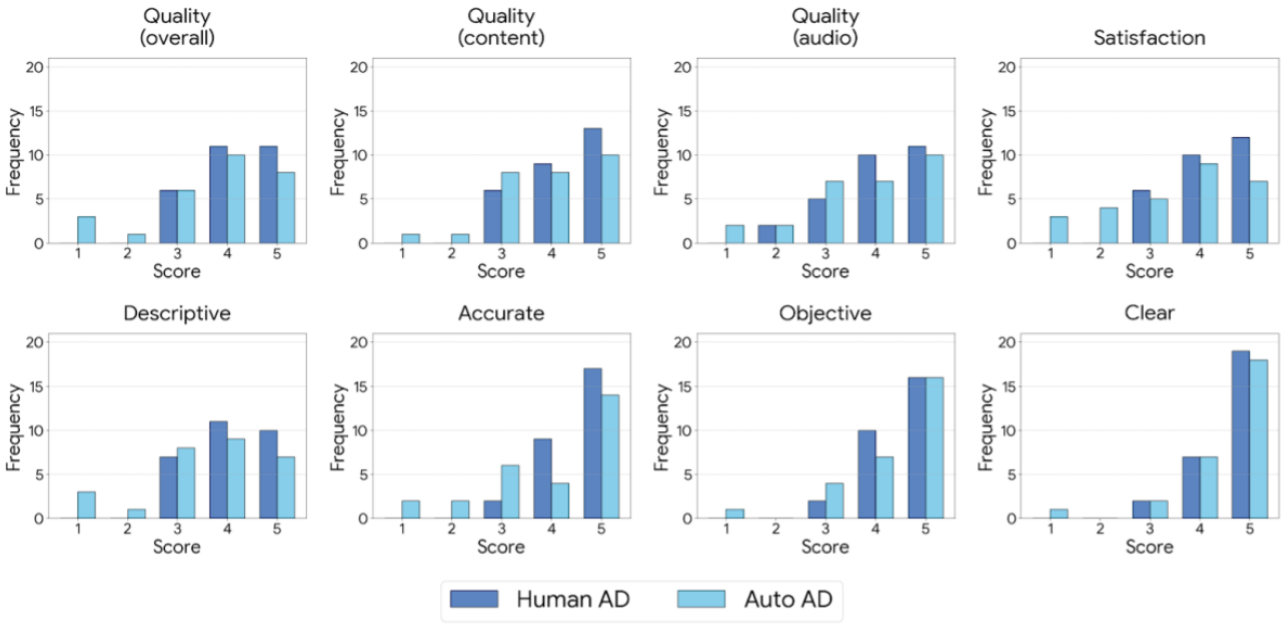


Figure 12. **BLV study.** Distribution of Likert scores across 8 key dimensions. BLV participants evaluated videos played with human written ADs and AutoAD (Gemini-2.5 Flash with Gd. + Timing ASR). They rated both AD versions consistently high (ratings of 4 or 5) across all metrics. A few of the videos with automated AD was rated lower with respect to audio quality, and in terms of overall quality and satisfaction when compared to the same videos played with human generated AD.

Questions for Sighted Raters

1. **Coverage:** The descriptions covered the important visual information necessary for a blind user to understand the video.
   - ☐ Strongly disagree. All important information was not covered.
   - ☐ Slightly disagree. Most important information was not covered.
   - ☐ Neither agree or disagree. Some important information was covered and some important information was missing.
   - ☐ Slightly agree. Most important information was covered.
   - ☐ Strongly agree. All important information was covered.
2. **Accuracy:** The descriptions were accurate.
   - ☐ Strongly disagree. All of the information was inaccurate (potentially misleading).
   - ☐ Slightly disagree. Most information was inaccurate.
   - ☐ Neither agree or disagree. Some information was accurate and some was inaccurate.
   - ☐ Slightly agree. Most information was accurate.
   - ☐ Strongly agree. All information was accurate.
3. **Relevance:** The descriptions included relevant information.
   - ☐ Strongly disagree. All of the information was irrelevant (not needed for a blind user to follow the video).
   - ☐ Slightly disagree. Most information was irrelevant.
   - ☐ Neither agree or disagree. Some information was relevant and some was irrelevant.
   - ☐ Slightly agree. Most information was relevant.
   - ☐ Strongly agree. All information was relevant.
4. **Consistency/Redundancy:** The descriptions account for information user already has from the video's audio and prior descriptions.
   - ☐ Strongly disagree. All descriptions repeated information the user would already know from the video's audio or prior descriptions.
   - ☐ Slightly disagree. Most descriptions repeated information the user would already know from the video's audio or prior descriptions.
   - ☐ Neither agree or disagree. Some of the descriptions provided new information but some of them included information

the user would already know from the audio or prior descriptions.
- ☐ Slightly agree. Most descriptions provided new information but a few descriptions included information the user would already know from the video's audio or prior descriptions.
- ☐ Strongly agree. All descriptions provided new information that supplemented the existing video's audio and prior descriptions.

5. **Overall Quality:** Overall, the content of the descriptions was high quality and engaging to listen to without the visuals.
   - ☐ Strongly disagree. All of the content of the narration had significant quality issues such that it was difficult to follow or not engaging to listen to. For example, the narration was overly repetitive and dull, or it had inconsistent references to characters, objects, or settings that made it difficult to follow.
   - ☐ Slightly disagree. Most of the content of the narration had significant quality issues such that it was difficult to follow or not engaging to listen to.
   - ☐ Neither agree or disagree. Some of the content of the narration was easy to follow and engaging to listen to. Some moments had quality issues that made the narration difficult to follow or not engaging to listen to.
   - ☐ Slightly agree. Most of the content of the narration was easy to follow and engaging to listen to. The language of the narration was mostly clear, illustrative, and engaging.
   - ☐ Strongly agree. All of the content of the narration was easy to follow and engaging to listen to. The language of the narration was entirely clear, illustrative, and engaging.
6. **Clarity:** The descriptions were clear and easy to follow considering the audio alone.
   - ☐ Strongly disagree. All descriptions were unclear or hard to follow when considering the information in the audio alone (e.g., used unclear or inconsistent references to characters, settings, or objects).
   - ☐ Slightly disagree. Most descriptions were unclear or hard to follow, but some descriptions were easy to follow when considering the information in the audio alone.
   - ☐ Neither agree or disagree. Some of the descriptions were clear and easy to follow, and some were not clear or hard to follow when considering the information in the audio alone.
   - ☐ Slightly agree. Most of the descriptions were clear and easy to follow when considering the information in the audio alone.
   - ☐ Strongly agree. All descriptions were clear and easy to follow from the audio alone.
7. **Descriptive:** The descriptions were descriptive in that they provided vivid details about objects, people, and settings while maintaining a concise narrative flow.
   - ☐ Strongly disagree. All of the descriptions were not descriptive. Instead, they were dull or repetitive, providing little useful information.
   - ☐ Slightly disagree. Most of the descriptions were not descriptive, but instead they were dull or repetitive.
   - ☐ Neither agree or disagree. Some of the descriptions included vivid details and some were dull or repetitive.
   - ☐ Slightly agree. Most of the descriptions included vivid details and the description mostly maintained a concise narrative flow.
   - ☐ Strongly agree. All of the descriptions included vivid details while maintaining a concise narrative flow.
8. **Objective:** The descriptions were objective in that they report what is visible without adding personal opinions or assumptions (e.g., about character's relationships, or identity attributes).
   - ☐ Strongly disagree. All of the descriptions contained subjective opinions or assumptions.
   - ☐ Slightly disagree. Most of the descriptions included subjective opinions or assumptions.
   - ☐ Neither agree or disagree. Some of the descriptions were objective but some of the descriptions included subjective opinions or assumptions.
   - ☐ Slightly agree. Most of the descriptions were objective but some were subjective.
   - ☐ Strongly agree. All of the descriptions were objective.
9. **Timeliness:** The descriptions appeared close to the time the event occurred in the video (usually within 3-5s).
   - ☐ Strongly disagree. All events were not described close to the time they occurred.
   - ☐ Slightly disagree. Most events were not described close to the time they occurred.
   - ☐ Neither agree or disagree. Some events were described close to the time they occurred.
   - ☐ Slightly agree. Most events were described close to the time they occurred.
   - ☐ Strongly agree. All events were described close to the time they occurred.
10. **Audio Issues:** All descriptions were clear to listen to and did not have audio issues such as overlapping speech.
    - ☐ Strongly disagree. All descriptions had audio issues such as overlapping speech. It is difficult to listen to the descriptions due to audio issues.

- ☐ Slightly disagree. Most descriptions had audio issues such as overlapping speech.
- ☐ Neither agree or disagree. There were some descriptions that were clear and some that would be difficult to listen to due to audio issues.
- ☐ Slightly agree. Most descriptions were clear without audio issues.
- ☐ Strongly agree. All descriptions were clear without audio issues.

11. **Gaps of Silence:** There were not large gaps of silence in the descriptions (6-10s quiet gaps).
    - ☐ Strongly disagree. All gaps of silence in the video did not have descriptions.
    - ☐ Slightly disagree. There were frequently long gaps of silence throughout the video.
    - ☐ Neither agree or disagree. There were some long gaps of silence throughout the video.
    - ☐ Slightly agree. There were long gaps of silence once or twice in the video.
    - ☐ Strongly agree. There were no long gaps of silence.

Prompts for AD Generation

Visual Information Prompt

```
Please watch this video and list out all the important visual events that happen in
the video narrative with timestamps.  Please be very descriptive with time stamps.

  The output format must be exactly like this:

 * One scene description per line.

 * Each description follows the format <timestamp> <description>
    • * The timestamp must be in the format 00:00:00.000 (hh:mm:ss.ms) followed by a
      "-".

 * The descriptions must be sorted by timestamp in ascending order.

 * **CRUCIAL**:  The descriptions must follow this format and cover the entire video.

Example output:
00:00:00.000 visual_event_description1
00:00:07.123 visual_event_description2
00:00:23.745 visual_event_description3

  Afterwards please watch the video and list all the text on screen and when it
appeared on the screen.
Please list all people in the video if there are any and what we know about the
people.
Finally, please provide a summary of the video narrative.
```

Figure 13. Prompt used to generate detailed visual information in the video. The responses are used when evaluating the VLM and human generated descriptions.

Simple Prompt

```
You are a professional audio description expert.  Create inline audio descriptions for
the provided video to describe visual events only when there is no speech or dialog.
For each audio description, provide the start time as a timestamp (e.g., 01:12:45.403
with HH:MM:SS.MS) then provide the text of the description.
```

Figure 14. Most basic baseline prompt used for generating audio description.

Guideline AD Prompt

```
You are an expert audio description writer.
Your task is to write high-quality audio descriptions for videos, ensuring
accessibility for visually impaired viewers.  Audio descriptions should narrate key
visual elements that are not already conveyed through the dialogue or other audio
elements.
Please adhere to the following guidelines:

 What to Describe:
* Describe what you see.  Focus on the essential visual information needed to
   understand the scene.
* Be concise.
* Always read on-screen text exactly as it appears.
* Be factual, objective, and avoid speculation.
* Use proper terminology and names whenever possible.
* All of the descriptions will need to fit in the space between dialog (assume 3-10
   seconds per word when spoken aloud).  If there is no space, then you need to skip
   the description.
* Try to match the mood of the video with your descriptions.
* Describe color only when it is vital to the comprehension of content.
** CRUCIAL: ** Your descriptions must cover the entire video.

 What NOT to Describe:
* Do not talk over the dialog or any other essential audio.
* Do not describe what can be inferred from the audio.
* Do not over-describe; less is more.  Keep descriptions brief and to the point.
* Do not interpret or editorialize.  Stick to describing what is visually present.
* Do not give away secrets, surprises, or sight gags before they happen.
* Do not censor (sex, violence, gore, emotions), describe them accurately.
* Do not describe obvious sound cues such as dialogue, a phone ringing, or a dog
   barking.

 Remember to only place descriptions between gaps of speech.  Strive for clear,
concise, and informative audio descriptions that enhance the viewing experience
without distracting from the original content.
The output format must be a valid JSON with a single field called "response"
containing the following:

* One scene description per line.

* Each description follows the format <timestamp>-STANDARD <description>
   • * The timestamp must be in seconds in the format 0:00:00.000 (h:mm:ss.ms)
     followed by a "-"
   • * The type of description should always be STANDARD
   • * The description must be in a single line.

* The descriptions must be sorted by timestamp in ascending order.

* **CRUCIAL**:  The descriptions must follow this format and cover the entire video.

* **CRUCIAL**:  The format must be valid.
```

Figure 15. Prompt used to generate audio description with just the guidelines and no video transcripts.

Guideline AD and ASR Prompt

```
You are an expert audio description writer.
Your task is to write high-quality audio descriptions for videos, ensuring
accessibility for visually impaired viewers.  Audio descriptions should narrate key
visual elements that are not already conveyed through the dialogue or other audio
elements.
```

```
The video transcript with timestamps (start and end time in seconds) is provided
below.  Make sure to generate descriptions only in between dialogs where there is
no speech in the video.  The descriptions should fit in the gaps when spoken out loud.
(assume you can fit in 1-3 words per second)

  <asr_transcript>
ASR_TRANSCRIPT_GOES_HERE_IF_AVAILABLE
</asr_transcript>

  Please adhere to the following guidelines:

  What to Describe:
 * Describe what you see.  Focus on the essential visual information needed to
    understand the scene.
 * Be concise.
 * Always read on-screen text exactly as it appears.
 * Be factual, objective, and avoid speculation.
 * Use proper terminology and names whenever possible.
 * All of the descriptions will need to fit in the space between dialog (assume 3-10
    seconds per word when spoken aloud).  If there is no space, then you need to skip
    the description.
 * Try to match the mood of the video with your descriptions.
 * Describe color only when it is vital to the comprehension of content.
 ** CRUCIAL: ** Your descriptions must cover the entire video.
What NOT to Describe:
 * Do not talk over the dialog or any other essential audio.
 * Do not describe what can be inferred from the audio.
 * Do not over-describe; less is more.  Keep descriptions brief and to the point.
 * Do not interpret or editorialize.  Stick to describing what is visually present.
 * Do not give away secrets, surprises, or sight gags before they happen.
 * Do not censor (sex, violence, gore, emotions), describe them accurately.
 * Do not describe obvious sound cues such as dialogue, a phone ringing, or a dog
    barking.

  Remember to only place descriptions between gaps of speech.  Strive for clear,
concise, and informative audio descriptions that enhance the viewing experience
without distracting from the original content.
The output format must be a valid JSON with a single field called "response"
containing the following:

 * One scene description per line.

 * Each description follows the format <timestamp>-STANDARD <description>
    • * The timestamp must be in seconds in the format 0:00:00.000 (h:mm:ss.ms)
      followed by a "-"
    • * The type of description should always be STANDARD
    • * The description must be in a single line.

 * The descriptions must be sorted by timestamp in ascending order.

 * **CRUCIAL**:  The descriptions must follow this format and cover the entire video.

 * **CRUCIAL**:  The format must be valid.
```

Figure 16. Prompt used to generate audio description with the audio transcripts and the guidelines.

Guideline + Timing ASR Prompt

```
You are an expert audio description writer.
Your task is to write high-quality audio descriptions for videos, ensuring
accessibility for visually impaired viewers.  Audio descriptions should narrate key
visual elements that are not already conveyed through the dialogue or other audio
elements.
The video transcript with timestamps (start and end time in seconds) is provided
below.  Make sure to generate descriptions only in between dialogs where there is
no speech in the video.  The descriptions should fit in the gaps when spoken out loud.
(assume you can fit in 1-3 words per second)

  <asr_transcript>
ASR_TRANSCRIPT_GOES_HERE_IF_AVAILABLE
</asr_transcript>

  Please adhere to the following guidelines:
paragraph **CRUCIAL TIMING & LENGTH RULES TO THINK AND FOLLOW:** 1.  **No Overlap:**
Descriptions MUST NOT overlap with any spoken dialogue in the provided transcript.
Use the timestamps to find silent gaps.  2.  **Calculate Gap Duration:** For each gap,
calculate the available time in seconds.

  What to Describe:
 * Describe what you see.  Focus on the essential visual information needed to
    understand the scene.
 * Be concise.
 * Always read on-screen text exactly as it appears.
 * Be factual, objective, and avoid speculation.
 * Use proper terminology and names whenever possible.
 * All of the descriptions will need to fit in the space between dialog (assume 3-10
    seconds per word when spoken aloud).  If there is no space, then you need to skip
    the description.
 * Try to match the mood of the video with your descriptions.
 * Describe color only when it is vital to the comprehension of content.
 ** CRUCIAL: ** Your descriptions must cover the entire video.
What NOT to Describe:
 * Do not talk over the dialog or any other essential audio.
 * Do not describe what can be inferred from the audio.
 * Do not over-describe; less is more.  Keep descriptions brief and to the point.
 * Do not interpret or editorialize.  Stick to describing what is visually present.
 * Do not give away secrets, surprises, or sight gags before they happen.
 * Do not censor (sex, violence, gore, emotions), describe them accurately.
 * Do not describe obvious sound cues such as dialogue, a phone ringing, or a dog
    barking.

  Remember to only place descriptions between gaps of speech.  Strive for clear,
concise, and informative audio descriptions that enhance the viewing experience
without distracting from the original content.
The output format must be a valid JSON with a single field called "response"
containing the following:

 * One scene description per line.

 * Each description follows the format <timestamp>-STANDARD <description>
    • * The timestamp must be in seconds in the format 0:00:00.000 (h:mm:ss.ms)
      followed by a "-"
    • * The type of description should always be STANDARD
    • * The description must be in a single line.

 * The descriptions must be sorted by timestamp in ascending order.

 * **CRUCIAL**:  The descriptions must follow this format and cover the entire video.

 * **CRUCIAL**:  The format must be valid.
```

Figure 17. Prompt used to generate audio description with the audio transcripts and the guidelines.

Prompts for AD Evaluation Metrics

Prioritization Prompt

```
You are evaluating if audio descriptions (AD) are covering visual information
necessary to understand the video.  For each line in the first AD output, score if
it is covered in the visual information (0 - not covered, 1 - fully covered).
Then, output only an array of all the scores like [0,0.5,1.0,1.0,0]

  First AD Output:
<ad_output>

  Second visual information of the video:
<dense_visual_info>
```

Figure 18. Prompt used to evaluate the prioritization of important events in the audio description based on detailed visual information.

On-Screen Text Narration Prompt

```
I will provide the on-screen text that occurs in the video.  For each on-screen text
segment, check if the text is reported in either the transcript or audio description
at a nearby time.  If a text chunk is not said in the transcript or audio description,
count the number of unsaid words in the chunk.  Then, return a list of unsaid word
counts e.g., [0, 0, 1, 0, 5, 0].  ONLY RETURN THE LIST.

  Here is the interleaved transcript labeled [TRANSCRIPT] and audio descriptions
labeled [AD]:
<interleaved_transcript_ad_str>

  Here is everything visual that occurred in the video including ON-SCREEN TEXT:
<visual_info_str>
```

Figure 19. Prompt used to evaluate whether on-screen text is narrated in the audio description.

Redundancy Prompt

```
I will provide the video transcript [TRANSCRIPT] and audio description [AD]
interleaved, such that they are in order as they would be played in the video.
Then, for each audio description [AD], check if the audio description repeats the
nearby transcript [TRANSCRIPT] or would otherwise be obvious from the audio alone
(e.g., describes offscreen door slam).  Score each AD between 0 and 1 where no repeat
- 0, slight repeat - 0.5, and egregious repeat - 1.0.  The purpose is to remove
unnecessary redundancy with the audio track from the audio description.  Then, return
a list of these assigned numbers e.g., [0, 0, 1.0, 0, 0, 0.5, 0.6].  ONLY RETURN THE
LIST.

  Here is the video transcript and audio description interleaved:
<interleaved_transcript_ad_str>
```

Figure 20. Prompt used to evaluate the redundancy of audio descriptions with the original transcript.

Timeliness Prompt (Visual Information-based)

```
For each audio description item, identify the corresponding visual information if one
exists, and report the time difference (in secs):

  [audio description time] audio description
[visual information time] visual information
**[audio description time - visual info time]**
The response should not contain any thing else.
For example:  -----

  [00:03:01.183] Ants crawl on banana pieces.
[00:03:12.257] Ants crawl on pieces of banana on rocks.
**[11]**

  [199] A light brown cockroach hangs from a screen, shedding its exoskeleton.
No audio description.

  **[N/A]**

  -----

  Here are the audio descriptions and visual information:

  Audio descriptions (timestamp isin hh:mm:ss.ms when the audio description starts in
the video):

  <audio_descriptions>
Visual information (timestamp isin hh:mm:ss.ms indicate when the visual appeared in
the video):

  <visual_info>
```

Figure 21. Prompt for evaluating the timeliness of audio descriptions by comparing them to pre-extracted visual information.

VideoA11y Prompt

```
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:

GROUND_TRUTH_DESC_GOES_HERE_IF_AVAILABLE
</ground_truth_desc>  <ground_truth_desc>
GROUND_TRUTH_DESC_GOES_HERE_IF_AVAILABLE
</ground_truth_desc>

  Model-generated video description:
GENERATED_DESC_GOES_HERE_IF_AVAILABLE
</generated_desc>  <generated_desc>
GENERATED_DESC_GOES_HERE_IF_AVAILABLE
</generated_desc>

  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"
```

Figure 22. Prompt used to evaluate the VideoA11y based metrics.