Metrics for Fine-Grained Evaluation of Inline Audio Descriptions

Subhashini Venugopalan[1] Tyler Roper[1] Jenny Tan[1] Jimmy Tobin [1] Emily Wilson[1]
Alicia Martin[1] Anton Kast[1] Amy Pavel[1,2]
[1]Google Research [2]UC Berkeley

Abstract

Audio descriptions (AD) that describe visual content in audio make videos accessible to people who can not see them. While humans create audio descriptions for film and television, audio descriptions remain largely unavailable for the vast majority of user-generated videos online. Now, publicly accessible Multimodal Large Language Models (MLLMs) can create audio descriptions on-demand. However, the quality of such descriptions remains unknown, especially for user-generated videos. We propose fine-grained evaluation metrics for assessing the quality of audio descriptions derived from audio description expert guidelines and a user study. We collect expert human descriptions for 400 user-generated videos, and then use our proposed metrics to compare the quality of MLLM-generated to human-created descriptions. While we find MLLM generated AD to accurately describe visual events, we find significant gaps in terms of audio quality. We synthesize these remaining gaps between human and MLLM-generated ADs using our metrics and qualitative analysis.

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 [10]. 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 [10, 11]. Audio descriptions (AD), or narrations of important visual content that avoid overlap with the video audio make videos accessible to BLV audience members [1, 17]. But manually created audio descriptions that are present on film and TV are rarely present for user-generated videos [10]. A long history of prior work across Computer Vision and Human Computer Interactions has thus explored semi-automated [3, 11, 13, 20] and automated approaches [4, 5, 8, 9, 18, 19]. Recent progress in multimodal large language models (MLLMs) has made automated descriptions increasingly accurate, but ***are these audio descriptions relevant, understandable, and enjoyable for BLV audience members?***

Prior work in Computer Vision and Human Computer Interaction has evaluated audio descriptions via comparison to ground truth [5], human evaluation with hired annotators or intended audience members [8], and more recently large model based evaluation [5, 8]. However, audience member evaluation is difficult to scale and high quality ground truth descriptions are rare for user-generated videos such that prior work used synthesized descriptions [5] or novice-created descriptions [8] as proxies. Course metrics for human- or model-based evaluation (*e.g.*, quality, satisfaction) have started to saturate on such proxies [8], and it can be difficult to assess how to improve user experience based on course metrics alone. Lower-level metrics can evaluate accuracy (*e.g.*, accuracy of character names [5]) and coverage (*e.g.*, similarity with ground truth [5]) but do not yet capture aspects of the user experience.

In this work, we propose a framework for automated evaluations of audio descriptions for user-generated videos. Our key idea is to use expert audio description guidelines to inform actionable automated metrics for AD that capture discrepancies between generated descriptions and known best practices. Towards this goal, we surface a general-purpose set of audio description best-practices synthesized from expert audio description guidelines [1, 2, 14, 15, 17] and create a comprehensive metrics based on these guidelines. Then, we collect a dataset of ground truth descriptions for user-generated videos crafted by teams of blind and sighted professional audio describers. Finally, we provide a comparison of expert-crafted audio descriptions with audio descriptions generated in a single step by a large model that takes video as input with several prompt variations to surface common problems and best practices.

2. Development of Metrics

To create an evaluation framework for audio descriptions that provides interpretable and actionable results for model improvement, we reviewed expert guidelines, then prioritized the most impactful guidelines to create metrics.

2.1. Expert-Informed Guidelines

We reviewed a set of expert audio description guidelines from 5 sources selected for comprehensiveness and diversity (American Council of the Blind's guidelines [1], DCMP's Description Key [2], ADLab's guide for describing film [14], W3C's Web Content Accessibility Guidelines [16], and YouDescribe's tips for description [15]), similar to prior work [13]. We also reviewed prior literature on BLV audience members perceptions of audio descriptions to surface deviations between expert best practices and specific BLV audience member preferences [6, 7, 10, 12]. In collaboration with researchers in Human Computer Interaction (HCI), Accessibility, and Computer Vision (CV), we then synthesized general-purpose best practices for inline audio descriptions (guidelines G1-G10):

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 [7, 10, 12]).
- **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 [5]).
- **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 [7, 10]).

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 purpose of this analysis, 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**).

2.2. Metrics for Evaluation

We create metrics for the guidelines evaluated to be most important during our pilot studies. For most guidelines, we use a large model to score the generated audio description against the selected guidelines and/or a ground truth audio description. All metrics except for **G7** and **G8** use an MLLM for evaluation and we include full prompts for each metric in the Appendix.

For **G1.1** (prioritization), we compare the generated description to the ground truth description to assess how well the generated description covers the content that a human expert covered in their own description. Specifically, we provide the ground truth description and generated description to an MLLM and ask to assess if each description in the first AD is covered in the second AD (coverage - recall), and then vis versa (coverage - precision)(Appendix Fig. 7).

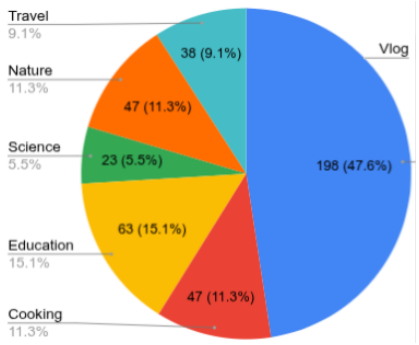
For **G1.2**, **G1.3**, and **G2**, we use an MLLM to score each individual audio description according to the guideline. For **G1.2** (redundancy), we provided an automated speech recognition (ASR) transcript interleaved with the generated AD (Appendix Fig. **??**) in order to an LLM with a prompt to assess how redundant each AD line was with the surrounding transcript on a scale from 0 (not redundant) to 1 (fully redundant). We then calcuated the mean of the values to achieve an overall redundancy score (Appendix Fig. 10). For **G1.3** (on-screen text), we first provided the video to an MLLM with a prompt to provide timestamps and describe all of the visual content in the video, all of the on screen text, and all of the characters to create a list of all **visual information** (Appendix Fig. 3). Then, we provide the AD and visual information to an MLLM with a prompt to assess the number of words in each on-screen text snippet covered by AD (Appendix Fig. 8). We calculate the mean number of missing words for all on-screen text snippets to assess coverage of on-screen text. For **G2** (timeliness), we provide 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 (Ap-

pendix Fig. 9). We calculate the mean and max absolute offset to achieve a timeliness score.

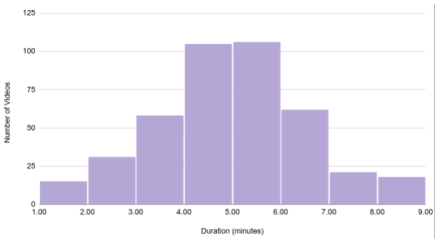
For **G7** and **G8**, we estimate the time ranges for each spoken audio description using an average speaking rate (200 words per minute) and assess the time ranges for speech in the video with an ASR transcript. We compute total seconds of overlapping audio descriptions with speech (**G7**, overlaps) and we compute the coverage of the audio time in both seconds of descriptions per silent video second and number of words per silent video segment (**G8**, audio coverage).

3. Dataset

The candidate videos for annotation were selected from YouTube. Candidates were chosen based on first identifying channels that had a fair number of subscribers (at least 1000 or more), and then selecting videos that were primarily in the 3-7 minute range, sometimes shorter or longer. We aimed to select videos that had at least some amount of silence, so that there is scope for including some description between the narration. We tried not to select more than 3-5 videos from the same channel. We obtained professional audio descriptions for the videos, where professionals filtered the videos and then annotated them with AD. The final dataset consists of 438 videos belonging to different categories as shown in Fig. 1a The videos are on average about 5 minutes in duration (see Fig. 1b). We describe the both the annotation process and overall video selection below.



(a) Categories of videos in the dataset.



(b) Duration of video clips.

Figure 1. Distribution of video types and duration.

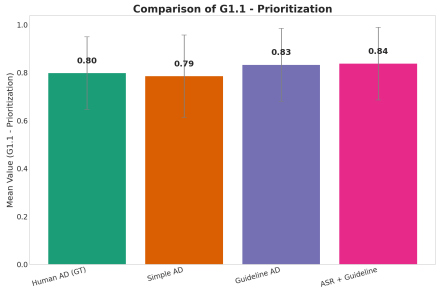
3.1. Professionally Annotated AD

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.

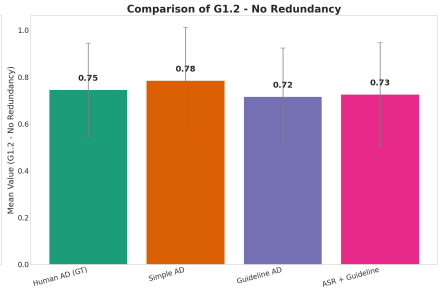
The **video selection process** for annotation began with a full review of each candidate video to assess its viability for audio description. A video was deemed non-viable if it contained nonstop dialogue or narration, leaving no temporal gaps for new descriptions, or if it was a silent vlog where existing 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. This curation strategy was applied to a total of 557 videos. 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.

3.2. Model Generated AD

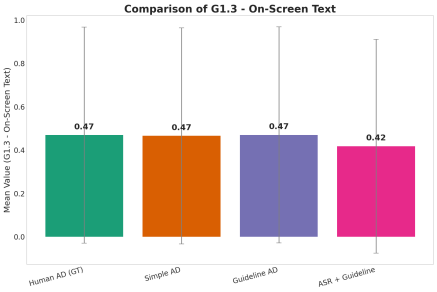
We also obtained audio descriptions from closed source multimodal model, primarily Gemini 2.5, which processes videos. We used 3 different prompting strategies, ranging from simple to more comprehensive to capture a range of descriptions. The **Simple Prompt** (in Appendix Fig. 4) very briefly instructs the model to generate inline audio descriptions for a given video and specifies the output format. A second prompt (**Guideline AD**, Appendix Fig. 5) provides a more detailed description of what is expected in an audio description, including several points from the guidelines for AD. The final most comprehensive prompt (**ASR + Guideline** prompt shown in Fig. 6) builds on the Guideline AD prompt by also including the transcript for the speech in the original videos.



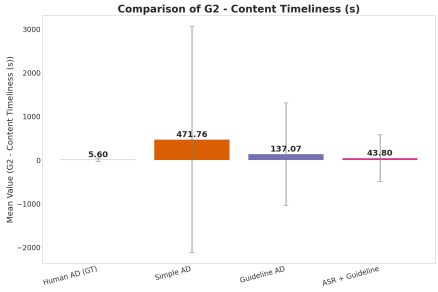
(a) Prioritization: AD captures visual info.



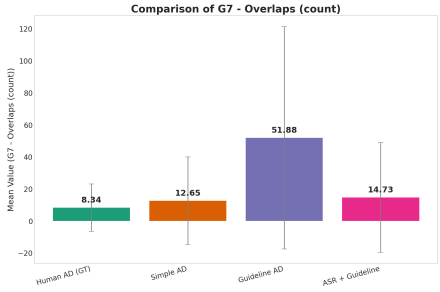
(b) No redundancy: b/w dialog and description.



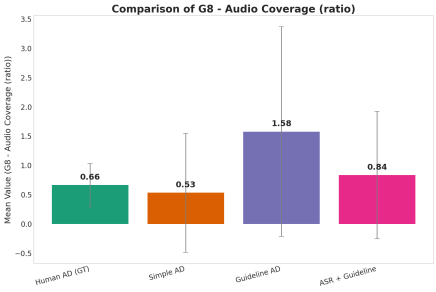
(c) All on screen text is captured.



(d) Timeliness: b/w visual event and description.



(e) Overlap counts between AD and dialog.



(f) AD covers the silence well.

Figure 2. Comparing AD from different systems on the 6 proposed metrics.

4. Experiments and Results

For each metric we compare the performance of human descriptions and model descriptions on 355 videos and report the results in Table 1.

The results highlight a clear trade-off between different prompting strategies for generating Audio Descriptions (AD). While all prompts demonstrates the strong performance in Prioritization (**G1.1**) and capture relevant visual events and do not appear to exhibit too much redundancy (**G1.2**), it is still hard for them to cover all on-screen text (**G1.3**). The difference between the prompts is much more clear when measuring timeliness (**G2**) and amount of overlap (**G7**). For instance, both *Simple Prompt* and *Guideline AD* prompt's effectiveness are severely undermined by a very high latency in Content Timeliness (**G2**), averaging over 471s and 137s respectively. Combined with the high overlap count of 51.88s, and higher than 1 audio coverage indicate that Guideline AD generates far more descriptions than the available silence which would affect the listening experience. In contrast, the **ASR + Guideline prompt achieves a much better balance**; with AD's that overlap less and are more timely ($43.80s$ vs Human at $5.80s$) and performs well on prioritization, suggesting that providing the model with ASR transcripts is critical for synchronizing descriptions with on-screen events. Overall, with timeliness and overlap counts still very far from Human Ground Truth annotations, the metrics provide a useful signal on the room for improvement.

5. Conclusion and Future Work

We present metrics for assessing audio descriptions inspired by expert guidelines then use these metrics to compare expert-crafted audio descriptions with model-crafted descriptions across several prompts. While our work identified remaining issues in AD (*e.g.*, overlap with transcript compared to ground truth descriptions), it also revealed places where in practice, expert AD conflicts with formal guidelines to fit the most important information in the time provided (*e.g.*, does not always describe on-screen text). We used our metrics to assess the relative performance of multiple descriptions, but in the future additional work will be required to assess the absolute performance of metrics in terms of accuracy, reliability, and correlation with human preferences. We also only created metrics for a subset of our guidelines related to errors that BLV audience members noticed the most (*e.g.*, an AD placed out of context much later in time). But, in the future we will create metrics for all guidelines. We used metrics as a way to evaluate generated descriptions, but in the future such metrics might be used to provide AD coverage information to audience members before they watch (*e.g.*, letting users know on-screen text is frequently missing in a recipe video), similar to metrics for video accessibility in prior work [10]. We hope that our work catalyzes future work in evaluating automated audio descriptions and ultimately improving audio descriptions for everyone.

References

[1] American Council of the Blind. American council of the blind, audio description project, guidelines for audio describers. https://www.acb.org/adp/guidelines.html, 2020. 1, 2

[2] DCMP. Description key. https://dcmp.org/learn/captioningkey/624, 2020. 1, 2

[3] 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

[4] 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

[5] 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

[6] 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. 2

[7] 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. 2

[8] 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

[9] 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

[10] 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, 2, 4

[11] 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

[12] 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. 2

[13] 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

[14] A. Remael, N. Reviers, and G. Vercauteren. Pictures painted in words: Adlab audio description guidelines. https://dcmp.org/learn/captioningkey/624. 1, 2

[15] The Smith-Kettlewell Eye Research Institute. Youdescribe. https://youdescribe.org/, 2022. 1, 2

[16] W3C. G8: Providing a movie with extended audio descriptions. https://www.w3.org/TR/WCAG20-TECHS/G8.html. 2

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

[18] 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

[19] 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

[20] 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

Appendix

Main results

Metric	Human AD (GT)	Simple AD	Guideline AD	ASR + Guideline
G1.1 - Prioritization	0.80 ± 0.15	0.79 ± 0.17	0.83 ± 0.15	0.84 ± 0.15
G1.2 - No Redundancy	0.75 ± 0.20	0.78 ± 0.23	0.72 ± 0.21	0.73 ± 0.22
G1.3 - On-Screen Text	0.47 ± 0.50	0.47 ± 0.50	0.47 ± 0.50	0.42 ± 0.49
G2 - Content Timeliness (s)	5.60 ± 41.27	471.76 ± 2582.31	137.07 ± 1175.77	43.80 ± 534.68
G7 - Overlaps (count)	8.34 ± 14.77	12.65 ± 27.27	51.88 ± 69.36	14.73 ± 34.21
G8 - Audio Coverage (ratio)	0.66 ± 0.37	0.53 ± 1.02	1.58 ± 1.79	0.84 ± 1.09

Table 1. Comparing performance on the proposed metrics of the ground truth and the different prompts used to generate the AD.

Prompts

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 3. 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 4. 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 5. 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 6. Prompt used to generate audio description with the audio transcripts and the guidelines.

Coverage Prompt

```
For each line in the first AD output, score if it is covered in the second AD output
(0 - not covered, 100 - fully covered).
Then, output only an array of all the scores like [0,50,100,0,0]

  First AD Output:
<ad_output>

  Second AD Output:
<gt_output>
```

Figure 7. Prompt used to evaluate the coverage of one audio description by another.

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

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

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