How well can VLMs rate audio descriptions: A multi-dimensional quantitative assessment framework

ANONYMOUS AUTHOR(S)

Digital video is central to communication, education, and entertainment, but without audio description (AD), blind and low-vision users are excluded. While crowdsourced platforms and vision-language-models (VLMs) expand AD production, quality is rarely checked systematically. Existing evaluations rely on NLP metrics and short-clip guidelines, leaving questions about what constitutes quality for long-form content and how to assess it at scale. To address these questions, we first developed a multi-dimensional assessment framework for uninterrupted, full-length video, grounded in professional guidelines and refined by accessibility specialists. Second, we implemented a comprehensive methodological workflow, utilizing Item Response Theory (IRT), to assess the proficiency of VLM and human raters against expert-established ground truth. Findings suggest that VLMs can approximate ground-truth ratings and often achieve higher alignment than human raters; however, qualitative analysis reveals that VLMs lack the diagnostic value of human feedback. These insights underscore the potential of hybrid evaluation systems that leverage VLMs alongside human oversight, offering a path towards scalable AD quality control.

ACM Reference Format:
Anonymous Author(s). 2018. How well can VLMs rate audio descriptions: A multi-dimensional quantitative assessment framework. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 33 pages. https://doi.org/XXXXXXX.XXXXXXX

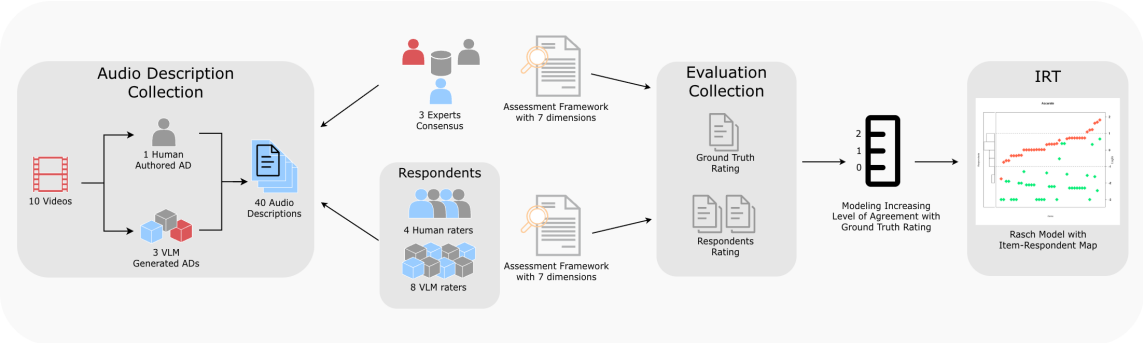


Fig. 1. Conceptual diagram of the major components in our evaluation workflow. From 10 videos, we collected 40 audio descriptions (1 human-authored and 3 VLM-generated per video). Three experts established ground-truth ratings using our seven-dimension assessment framework. Additional respondents, four humans and eight VLMs, applied the same framework, and their ratings were compared to the ground truth. Item Response Theory (IRT) was then used to model evaluator proficiency and item difficulty.

1 Introduction

Audio description (AD) narrates key visual information in videos, enabling blind and low-vision (BLV) audiences to access visual media that would otherwise be inaccessible. Professional AD enhances comprehension and inclusion, but producing it is resource-intensive, limiting coverage given the scale of digital content available today [29]. To expand access, volunteer-driven authoring tools have emerged to lower barriers to AD creation through streamlined interfaces

and collaborative editing [11, 31, 42]. More recently, automated approaches have aimed to further extend coverage by leveraging large language models (LLMs) and vision–language models (VLMs) to generate scene-aware descriptions [13, 36, 38]. While these approaches expand coverage, their quality is often inconsistent, which can negatively impact the viewing experiences of BLV users. Studies on the AD authoring tools such as Rescribe show that novices, when not aided by guidance, make timing and placement errors that reduce the usability of their descriptions [31]. Similarly, AI-based systems have been shown to misidentify characters or hallucinate details [9], produce verbose or repetitive narration that confuses rather than supports viewers [38], and achieve high benchmark scores while misaligning with human judgments of AD quality [32]. These limitations highlight that expanding AD coverage through volunteer or AI contributions must be accompanied by systematic quality checks, calling for a reliable mechanism to assess quality.

Existing AD quality assessments face two main limitations. First, they typically focus on isolated short clips, which fails to reflect real-word practice, whereas both volunteer and AI-generated AD now span uninterrupted full-length videos [21, 37, 44]. Second, they rely on user studies with BLV participants, crowdsourced workers, or professional experts [20]. While valuable, such studies are resource-intensive, and ratings from non-experts can be unreliable [39].

Our study addresses these gaps through two guiding research questions (RQs):

(1) **RQ1.** What constitutes AD quality in practice? How can professional guidelines and expert knowledge be operationalized into a systematic assessment framework?
(2) **RQ2.** Can VLMs approximate expert judgements and provide reliable support when human review is costly, building towards scalable evaluation of AD quality?

To address **RQ1**, we engaged consultants with extensive experience working with BLV users in assistive technology and multimedia accessibility to identify what matters in full-length AD beyond the established *content* dimensions found in professional guidelines [3]. Through this process, we identified a key gap: the absence of *formatting* dimensions, such as delivery method and timing, which capture how descriptions are integrated into the audiovisual flow and are essential to BLV audiences' viewing experience in practice. These insights were formulated into a multi-dimensional assessment framework designed for evaluating AD quality in full-length video. We also consulted with blind consultants including a blind professional AD quality controller and a blind scientist to further validate the framework.

To answer **RQ2**, we conducted a study that engaged both human accessibility experts and state-of-the-art VLMs as evaluators. First, we collected a corpus of volunteer-authored and AI-generated audio descriptions for full-length videos. Second, using our assessment framework, three experts first established ground-truth ratings for a set of volunteer-authored and AI-generated descriptions, with all evaluations conducted blind to source. Next, additional human respondents were recruited, and VLMs were prompted to complete the same evaluation task, allowing us to assess how their ratings aligned with the ground truth. Finally, we applied Item Response Theory (IRT), a statistical methodology from education and psychology, to analyze evaluator ability and description difficulty. This approach allows us to examine whether VLMs approximate expert ratings in a reliable and interpretable way. Our goal is to support AD evaluation that can scale alongside growing production demands, while preserving the validity of quality assessment.

Our paper makes two primary contributions.

First, we contribute a multi-dimensional assessment framework for evaluating audio description quality. While most current work evaluates AD using a single, one-dimensional scale based on textual similarity [5, 8, 30], recent research such as Li et al. has moved toward multi-dimensional evaluation by combining content dimensions with NLP metrics [20]. We extend this direction with a critical insight: although accessibility guidelines discuss formatting practices,

neither existing frameworks nor guidelines have formalized formatting into a measurable evaluation dimension. This gap is particularly significant for full-length video content in practical contexts. Our assessment framework addresses this by integrating established guidelines with expertise from accessibility experts and blind consultants, integrating five content dimensions and two critical formatting dimensions (timing and delivery). While subsequent work may refine the number and definition of dimensions, we demonstrate that formatting considerations are crucial and argue that the community should continue moving beyond one-dimensional constructs.

Second, we present a comprehensive methodological workflow for evaluating VLM and human performances on complex rating tasks, specifically the assessment of AD quality. This workflow adapts established principles from education and psychology to a novel context within Computer Science and Human-Computer Interaction (HCI). Developed by an interdisciplinary team, including experts in measurement, test construction and quality control, this approach yields more reliable results and reveals nuanced insights that simple accuracy metrics overlook. We apply this workflow, as illustrated in Fig. 1, to evaluate the proficiency of VLMs and humans as AD raters using our developed framework. Beyond this specific use case, the workflow is designed to be generalizable, offering the research community a practical and replicable methodology. This is a timely contribution as the field seeks robust tools to understand the growing capabilities and limitations of VLMs across diverse domains.

Our work is among the first in HCI and AI communities to focus on evaluating audio description quality on extends beyond short, isolated clips to full-length, uninterrupted videos through a combination of expert judgment and VLMs. The application of this workflow yields the following secondary contributions:

- Empirical findings from a mixed-methods study with accessibility experts, human raters, and state-of-the-art VLMs. Analyzed using Item Response Theory (IRT) to model rater proficiency and item difficulty on the same scale, these findings reveal where VLMs approximate expert judgments and where they diverge.
- Design implications for scalable, hybrid evaluation workflows that combine the efficiency of VLM ratings with the diagnostic value of human feedback, advancing both accessibility practice and the broader HCI agenda on human–AI collaboration in evaluation.

2 Related Work

2.1 Volunteer and AI-Generated AD

Audio description has been shown to improve comprehension, satisfaction, and engagement for BLV audiences [26, 34]. To support consistent practice, professional organizations have established standards for quality: the Described and Captioned Media Program (DCMP) [3] provides educational guidelines, while the National Center for Accessible Media (NCAM) at WGBH Educational Foundation, a Boston-based public media organization, established conventions for broadcast and film [27]. Commercial providers, such as 3Play Media, have incorporated AD into broader accessibility workflows with professional production processes [4]. Quality assurance typically involves multi-stage review: scripts are drafted and peer-reviewed, editorial checks are applied, recordings are tested for clarity and synchronization, and in some cases pilot testing with BLV users is conducted [29]. These processes make professional AD the gold standard for quality, but quality checks remain descriptive and guideline-driven, relying on expert oversight rather than quantitative assessment protocols.

To expand access beyond commercial and broadcast domains, crowdsourced platforms such as LiveDescribe, Rescribe, and YouDescribe adapt professional guidelines for non-experts [11, 31, 42]. While such community-driven efforts have made AD more widely available, their outputs are rarely reviewed against accessibility standards, resulting in variability

in accuracy, clarity, and pacing. Prior studies of volunteer describers confirm these challenges: without guidance, novices introduce timing and placement errors that reduce usability [31] and produce only 60% of the overall audio description quality compared to professionals, often with mismatched loudness, incomplete pacing, and mismatched placement strategies [25]. Beyond quality concerns, volunteer labor is strained by the explosive growth of video content on platforms like YouTube, TikTok, and Instagram, where uploads far exceed the capacity of manual description.

To address this bottleneck, researchers have explored automation. Early systems used multimodal narration [38], while recent pipelines leverage LLMs and VLMs for scene-aware descriptions [13, 36, 41]. While AI systems can generate descriptions quickly at scale, persistent quality problems remain. Automated descriptions may confuse or invent characters and objects[9, 32], deliver narration that is cluttered and at the wrong time, disrupting the viewing experience [38]. Moreover, despite strong scores on captioning benchmarks, their outputs diverge from human judgments of accessibility-critical qualities such as relevance and pacing [32].

The absence of quantitative assessment is especially stark for volunteer-authored and AI-generated AD, where growing volumes of content lack the rigorous expert-driven review of professional practice. Our work addresses this gap by introducing a comprehensive assessment framework to evaluate these emerging forms of AD at scale.

2.2 Efforts to Evaluate AD Quality

Evaluating the quality of audio descriptions is essential for ensuring their usability, particularly for BLV audiences who rely on AD for comprehension and engagement. Most evaluation efforts fall into two main tracks: automatic metrics and human-centered judgments.

Datasets such as VALOR [21], YouCook2 [44], and VATEX [37] support AI-generated AD research, with performance typically reported using NLP metrics like BLEU [30], METEOR [8], and CIDEr [14]. Content-based approaches like SPICE [5] extend this by parsing scene graphs to compare objects, attributes, and relationships semantically. While these metrics are efficient, objective, and reproducible, they were developed for general captioning tasks. As a result, they privilege surface similarity and fail to capture accessibility-critical qualities such as relevance and temporal alignment; factors that determine whether BLV audiences experience AD as supportive or disruptive.

For both AI-generated and volunteer-authored AD, evaluation typically relies on human participants. Studies with BLV users often adopt subjective methods, but no standardized user-based metric exists. Existing approaches often use overall quality ratings (e.g., Likert scales on understanding or satisfaction) [10], or task-based methods where BLV participants flag missing or confusing information [38]. While informative, these methods face persistent challenges: recruiting BLV participants remains difficult due to limited population size, reliance on advocacy organizations, accessibility logistics, and participant fatigue [2, 24]. Feedback from sighted participants, especially when gathered via crowdsourcing platforms like Amazon Mechanical Turk, also suffers from quality issues, including inattentive workers, bots, and inconsistent responses [39].

Recent work such as VideoA11y [20] marks an important step forward by combining accessibility-informed evaluation with standard NLP metrics. Their framework included four custom dimensions, *descriptive, objective, accurate, and clear*, alongside six standard captioning metrics, offering a more comprehensive view of AD quality than either approach alone. Their evaluation, conducted with 347 sighted participants, 40 BLV participants, and 7 professional describers, offered breadth and external validity but was resource-intensive and difficult to replicate. Their analyses were further limited to benchmark datasets such as VALOR, YouCook2, and VATEX, which contain only isolated 10–15 second clips.

We address these limitations by extending the scope of AD evaluation in three critical directions. First, we extend evaluation to full-length AD, ranging from 1:30 to 5:05 minutes. Each video represents complete, uninterrupted content,

where continuity and delivery are critical. Second, we introduce *formatting* dimensions, delivery method and timing, which no prior framework has addressed despite their importance to usability. Third, we explore whether VLMs can serve as evaluators, making AD quality assessment more sustainable and scalable by applying Item Response Theory (IRT) to model evaluator reliability and task difficulty.

2.3 LLM/VLM as Evaluators and the role of Item Response Theory

As research on automated AD generation grows, scholars have also begun to ask whether LLMs and VLMs can serve not only as generators but also as evaluators of content. In NLP, the LLM-as-a-Judge paradigm has been tested in tasks such as summarization, translation, and dialogue, where model-based rankings often correlate with human preferences [12, 43]. Multimodal extensions such as Prometheus-Vision demonstrate that VLMs can deliver fine-grained judgments of captions and visual explanations closely aligned with human ratings [19]. Within HCI, researchers frame models less as replacements and more as collaborators in evaluation workflows. EvalAssist, for example, integrates LLMs into interactive processes that reduce evaluator effort while preserving alignment with expert assessments [6], underscoring a growing interest in hybrid strategies that combine human expertise with machine scalability.

Despite momentum around model-as-judge frameworks, most evaluations still rely on simple validation measures—accuracy against majority vote, correlation with human scores, agreement rates, or win-rate comparisons [12, 19, 43]. While useful, these metrics collapse complex judgments into a single number, overlooking variation in rater reliability, prompt sensitivity, and task difficulty. This is especially limiting for accessibility tasks like evaluating AD, where nuanced qualities such as relevance, equality and pacing directly shape BLV comprehension.

We argue that Item Response Theory (IRT) offers a more robust methodology for modeling evaluation. Developed in educational testing, IRT treats *items* as test questions and *respondents* as students, jointly modeling item difficulty and respondent ability [1]. In our context, items are audio descriptions (volunteer-authored and AI-generated), while respondents are human raters and VLMs prompted to provide ratings. This framing enables us to model evaluation as an interaction between raters and tasks: not only whether a rater is capable in absolute terms, but how their ability varies with the difficulty of the description. In our case, IRT estimates which audio descriptions are harder to evaluate, and which raters, human or VLM, are more likely to align with expert ground-truth ratings, even allowing for partial agreement rather than exact matches.

We build upon these psychometric foundations to propose a comprehensive methodological workflow for evaluating AI raters. While IRT has already been adapted in HCI to handle heterogeneity in inspectors, improve questionnaire analysis, and calibrate subjective judgments [33, 35], and more recently in natural language generating evaluation to reveal systematic differences in human rater reliability [17, 18], these methods have not yet been integrated into a unified workflow for assessing hybrid human–AI evaluation ensembles. To the best of our knowledge, no prior work has applied this workflow in accessibility contexts, specifically the assessment of VLMs as evaluators of AD quality. By doing so, our study extends this psychometric method to a new domain, providing a richer account of both evaluator ability and AD difficulty, and advancing hybrid human–AI evaluation workflows in accessibility.

3 Assessment Framework

3.1 Accessibility Consultants Insights and Motivation

Our assessment framework was shaped through an iterative, multi-stage refinement process.

We began with the DCMP Description Key guidelines—the current gold standard in professional practice—whose five content principles have been developed through practitioner consensus. This provided our initial content foundation.

Next, we consulted an experienced professional describer who trains new describers and evaluates their work. She emphasized not only accurate depiction of visual elements and on-screen text, but also the importance of prioritization (avoiding under and over description), and two formatting considerations she routinely uses during evaluation: determining when to use inline versus extended description and assessing track placement. Her criteria highlighted that delivery choices are central to judging real-world AD quality.

We then consulted a specialist who provides training for Teachers of Students with Visual Impairments and has extensive expertise in assistive technology, universal design and digital multimedia accessibility. Drawing on her experience training community practitioners in accessibility, she observed that the emerging criteria naturally cluster into two overarching categories: *content* and *formatting*. This framing crystallized a key insight: description quality depends equally on what is described and how and when descriptions are delivered.

While prior research has largely focused on content quality, evaluating whether descriptions are accurate or descriptive enough, the consultants stressed that formatting choices often determine whether content is usable in practice. Formatting encompasses both *Track Placement* (timing descriptions relative to dialogue or sound) and the choice between *Inline* delivery, which inserts narration into natural pauses, and *Extended* delivery, which briefly pauses playback to deliver additional detail and resume the program. Poor timing can cause even accurate content to clash with dialogue or arrive too late, disrupting comprehension and reducing enjoyment for BLV users [26]. Similarly, the omission of extended narration in dialogue-heavy, fast-paced, or educational contexts can leave critical details unspoken, preventing audiences from fully understanding the content [4].

Despite its importance, formatting has been largely overlooked in AD research. Benchmarks such as YouCook2 [44], VATEX [37], and VALOR [21] contain only short clips, where delivery issues rarely surface, and most AI-based pipelines default to generating descriptions for silent sections, implicitly producing only inline-style narration [13, 41]. This narrow focus sidelines delivery concerns, even though volunteers on platforms like YouDescribe and Rescribe are now authoring full-length AD where extended narration and precise timing are essential.

As one consultant noted: *"audio description without careful formatting is like a paper with strong ideas but no structure, or conversely, polished formatting with no substance."* In both cases, the outcome loses its value. This perspective motivated us to design an evaluation model that treats content and formatting as equally critical dimensions for assessing AD quality.

3.2 Formalization of the Assessment Framework

Content Dimensions. The DCMP guidelines are widely recognized as the gold standard for audio description. Their *"Quality Keys"* identify five essential dimensions that descriptions must satisfy to be considered high quality [3]:

(1) **Accurate**: Descriptions are factually correct and error-free.
(2) **Prioritized**: Information critical for comprehension and enjoyment should be emphasized, while less important details are minimized.
(3) **Appropriate**: Language should be suited to the target audience, maintaining simplicity and conciseness.
(4) **Consistent**: Terminology, tone, and pacing should align with the program's style and remain uniform throughout the narration.

(5) **Equal**: Descriptions should preserve the program's meaning and intent for equitable access without distortion or bias.

Formatting Dimensions. As emphasized by our consultants, delivery choices, both method and track placement, are as critical as content in determining AD quality. With crowdsourced platforms now supporting both inline and extended narration, describers increasingly face decisions about how to deliver descriptions effectively. To account for this, we incorporated two delivery-focused dimensions into our evaluation model, drawing directly on NCAM broadcast guidelines [27], DCMP standards [3], and professional industry practices such as 3Play Media [4].

(6) **Strategic Use of Description Method (Inline vs. Extended)**: Professional guidelines recommend inline narration as the preferred method when natural pauses allow visual details to be conveyed without disrupting the program. Extended narration is advised when natural pauses are insufficient, for instance, in dialogue-heavy, text-heavy, noisy, or fast-cut videos where critical information would otherwise be lost.
(7) **Timing and Placement**: Guidelines emphasize inserting narration as close as possible to the relevant visual action while avoiding overlap with dialogue or essential sounds. Both NCAM and DCMP recommend using natural pauses, aligning narration with the visual timeline, and where appropriate, allowing pre-description if it clarifies the scene.

After formalizing the assessment framework, we sought additional perspective from two blind professionals who bring lived experience as BLV users alongside their expertise. A quality-control (QC) specialist who reviews professional AD confirmed that the framework captures the criteria she utilizes in her evaluations and contributed explicit rating-level labels drawn directly from her QC workflow. A scientist and inventor with decades of experience leading R&D teams in accessible technology development also affirmed the structure of the framework and emphasized the importance of the formatting dimensions, particularly the role of extended descriptions in making scientific or dialogue-dense content comprehensible.

The resulting assessment framework is grounded in professional standards and guided by insights from accessibility consultants, while addressing gaps in prior approaches. It extends prior approaches by incorporating two *formatting* dimensions, delivery method and timing, which consultants identified as central to usability but that have been overlooked in short-clip evaluations. By integrating both content and formatting, the framework provides a comprehensive basis for evaluating volunteer-authored and AI-generated AD and supports more systematic, practice-aligned assessment.

4 Datasets

4.1 Videos Selection

To create a stimulus set that was both varied in quality and genre, we adopted a three-step selection process.

First, we drew from a broad pool of videos containing volunteer-authored audio descriptions publicly available on the crowdsourced platform namely YouDescribe [42]. Each video is rated by the community on a 1-5 scale (5 is better); the videos we selected ranged from 2 to 5 stars, signaling variability in quality. Although we lacked information about the source and criteria of these star ratings, they served as an initial filter to capture diversity in quality levels.

Second, an accessibility consultant with experience working with BLV users and training new audio describers conducted a pre-screen of this pool. The consultant sorted videos into stronger or weaker AD based on overall impression. This sorting was not used in later analyses; instead, this step helped ensure our final sample captured a range of descriptive demands and perceived quality.

Entertainment

(a) Star Wars: The Rise of Skywalker – Teaser(2:04)

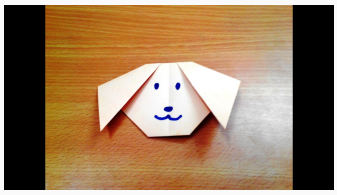
(b) Lady Bird | Official Trailer HD | A24 (2:34)

(c) Frozen Teaser (2013) - Disney Animated Movie (1:30)

(d) Elf Clip - Buddy Realizes He's Human (2003) (1:37)

How-to & Style

(e) 3 Ways to Make Homemade Pickles (2:23)



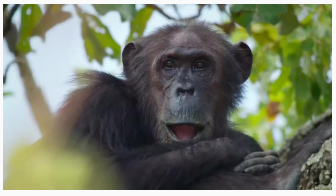
(f) How to Make an Origami Dog Face (2:47)

(g) Quick and Easy 5-Minute Makeup Tutorial (5:01)

Education

(h) Non-Newtonian Fluids: Crash Course Kids (4:20)

(i) Bald Eagle | Animals for Kids | All Things Animal TV (3:11)



(j) jane goodall (2:43)

Fig. 2. Ten YouTube videos used in the evaluation, grouped by category: Entertainment (top), How-to & Style (middle), and Education (bottom). Subcaptions show the YouTube titles with its video length

.

Third, drawing on the consultant's pre-screening, we further curated ten videos spanning multiple genres and audiences. Entertainment trailers (*Star Wars*, *Lady Bird*) required maintaining narrative continuity amid dense visuals and tonal shifts. Family-oriented films (*Elf*, *Frozen*) demanded age-appropriate language while conveying humor and action. How-to & Style clips (*3 Ways to Make Homemade Pickles*, *Quick and Easy 5-minute Makeup tutorial*, *How to Make and Origami Dog Face*) required stepwise clarity to describe fine-grained hand movements. Education videos ranged from the rapid-fire style of *Crash Course for Kids* to the child-oriented *Bald Eagle | Animals for Kids* and the slower-paced *Jane Goodall* at *Gombe*, which required integrating factual narration with scenic content.

By intentionally combining crowdsourced quality signals, expert pre-screening, and genre-based curation, we produced a heterogeneous yet purposeful stimulus set. The dataset reflects real-world demand from blind and low-vision

audiences while also providing a meaningful testbed for examining how both human and VLM raters assess AD quality across genres, audiences, and levels of descriptive difficulty.

4.2 Audio Description

For each of the ten videos, we prepared four audio description (AD) versions: one volunteer-authored and three generated by VLMs, resulting in 40 total audio descriptions.

4.2.1 Volunteer-authored Audio Description. Volunteer-authored recordings uploaded to the crowdsourced AD platform were transcribed with Whisper (≈98% accuracy). Transcriptions were then reviewed by our team, and any errors were manually corrected to preserve the authenticity of the volunteers' original audio descriptions. To anonymize speakers and standardize delivery, transcripts were resynthesized with a single synthetic voice using Google Cloud Text-to-Speech. Speech rates were adjusted to approximate the pacing of the original volunteer (e.g., faster for action-oriented trailers and slower for nature content) thereby preserving the intent of the narration while removing potential bias from vocal delivery.

4.2.2 VLM-generated Audio Description. Three state-of-the-art VLMs, **Qwen2.5-VL** [7], **Gemini 1.5 Pro** [16], and **GPT-4o** [28], produced alternative descriptions following a common generation pipeline. We selected these models based on their performance on the large-scale Video-MME benchmark [15], which evaluates more than 50 VLMs on multimodal reasoning and temporal understanding. At the time of system development, Gemini ranked first overall, Qwen-VL fifth, and GPT sixth. Together, these models represent top performers across proprietary (Gemini, GPT) and open-source (Qwen) families, while also being among the few high-ranking systems accessible for research. This selection allows our stimulus set to reflect the diversity of current VLM paradigms without sacrificing practical reproducibility.

Each video was segmented into coherent scenes based on visual similarity, aligned with dialogue transcripts, and then provided as input to the models. Prompts were adapted from professional AD guidelines and emphasized factual accuracy, conciseness, neutrality, and timing awareness while discouraging hallucination and over-description. Model outputs were further optimized to reduce redundancy and improve narrative coherence, then synthesized into speech using the same Text-To-Speech system as the volunteer-authored descriptions. All versions supported both inline and extended narration. This process created a dataset that captured both authentic community-authored descriptions and systematically generated alternatives, standardized for a fair comparison in evaluation.

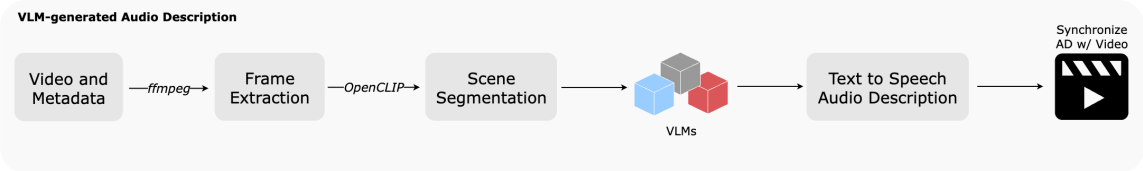


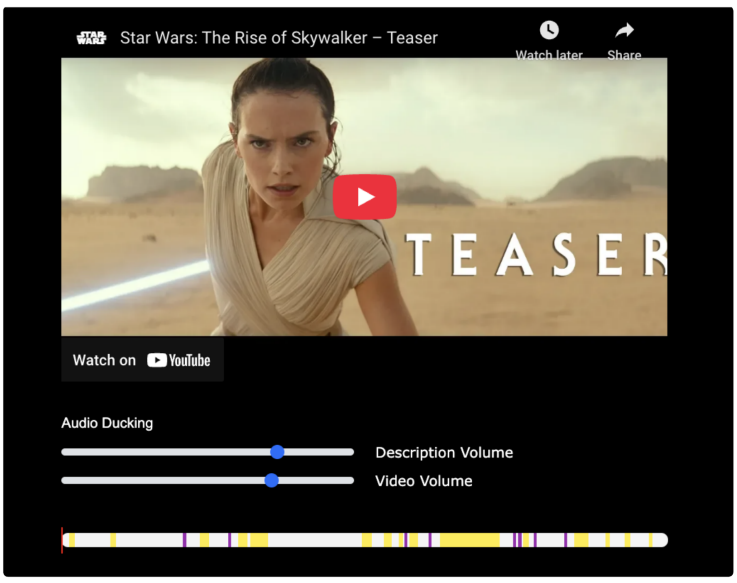
Fig. 3. VLM-generated description pipeline. Videos are segmented into scenes, a VLM generates scene-level descriptions, a second VLM pass optimizes them, and the outputs are resynthesized and synchronized with the video.

5 Study Design

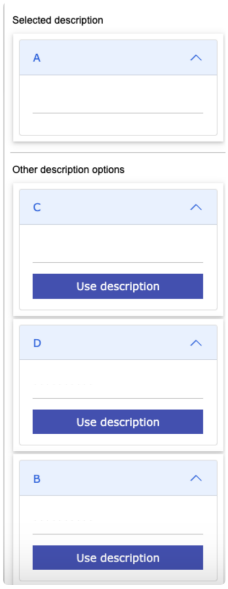
5.1 Establishing Ground Truth Ratings

We recruited three accessibility consultants with extensive experience supporting BLV communities, including mentoring novice describers, creating and reviewing accessible media resources, and training BLV users on assistive technologies. Using the assessment framework, the experts rated descriptions on a 1–5 ordinal scale across all seven dimensions, supported by comprehensive documentation including definitions, explanatory notes, illustrative examples, and evaluation criteria for each dimension (full materials are available in the Appendix). Their consensus ratings established the ground truth against which other respondents' ratings would be compared.

5.1.1 Interface. The experts interacted with a custom-built interface designed to present audio descriptions in context with the video (Fig.4). Inline narration segments appeared as yellow overlays embedded directly in the playback timeline, while extended narrations were shown as purple markers that briefly paused playback for longer insertions. The interface also included audio ducking controls to balance the relative loudness of narration with the original audio track. These features allowed raters to clearly perceive both the content of the description and its delivery format.



(a) Interface showing inline (yellow) and extended (purple) narration markers on the timeline, with audio ducking controls.



(b) Video player with four available AD versions.

Fig. 4. Viewing interface for synchronized playback of audio description with video.

5.1.2 Evaluation Procedure. For every video, four AD versions were available, one volunteer-authored and three generated by VLMs, displayed on the bottom right of the video player (Fig. 4b). To establish a *ground truth baseline*, accessibility experts independently evaluated these descriptions following a tightly controlled procedure. Anonymization and randomization were applied at multiple levels to minimize bias. First, all descriptions were relabeled with neutral identifiers (A–D), with the mapping randomized separately for each video. For example, in one video the label "A" might

correspond to the human-authored description, while in another video "A" could correspond to a Qwen-generated version. Similarly, "B" could denote a Gemini-generated description in one case but a GPT-generated description in another. Second, the order of videos were also randomized for each expert. Third, within each video, the four versions were presented in a randomized order so that no system consistently appeared earlier or later. Thus, one might see the sequence A–B–D–C for the first video and D–A–C–B for the second. These measures reduced recognition of description provenance and encouraged independent evaluation.

Each expert was provided with a personalized instruction sheet specifying their randomized sequence of videos and AD versions. For each version, participants followed a link that opened the interface with the current AD displayed under *"Selected description"* as seen in Fig 4b. Experts then listened to the AD and completed an evaluation form rating it across seven dimensions, with an optional text field for qualitative comments. To support progress tracking, the instruction sheet included a checklist of completed tasks.

To establish the ground truth for analysis, we applied a majority/consensus rule to these expert ratings. When two experts independently assigned the same score, that value was taken as the ground truth. In the less frequent cases where all three experts diverged, we resolved disagreement by selecting the median rating, ensuring that the reference score reflected central tendency rather than privileging any single rater. This process produced the expert reference ratings that serve as the ground truth in our study.

5.2 Human Respondents

We recruited four additional human respondents, who had exposure to accessibility-related work, such as supporting learners with disabilities, participating in accessibility initiatives, or working with inclusive educational materials. Because the task involved evaluating 40 audio descriptions across seven dimensions, we selected participants with enough familiarity to engage meaningfully with the assessment framework and the demands of the task. Their background enabled them to complete the evaluation reliably and provide an additional perspective on how the assessment framework is applied in practice.

They then followed the same evaluation process as the experts described in Section 5.1. For each video, they assessed four AD versions (one volunteer-authored and three VLM-generated) using the same interface, instructions, and seven-dimension rating form. The same anonymization and randomization procedures prevented ordering effects and ensured that respondents could not identify the source of the AD (human authored or AI-generated). Their evaluations were combined with VLM-generated ratings to model agreement with the expert-established ground truth.

5.3 VLM Respondents

We prompted three VLMs: Qwen2.5-VL, Gemini 1.5 Pro, and GPT-4o to be respondents. They evaluated the same audio descriptions using our multi-dimensional assessment framework described in Section 3.2. We asked VLMs to give ratings as well as justifications for their ratings to probe how the VLM interprets the assessment framework and how its reasoning aligns or diverges from human raters.

Each model evaluated all AD versions, including its own output, the AD generated by the other two VLMs, and the human-authored AD. This cross-evaluation design reduces the likelihood of models only "self-evaluating" and enables us to examine whether they apply the assessment criteria consistently across sources. These VLM ratings were analyzed alongside expert and additional human raters in the IRT model, providing a shared basis for comparing how each rater type applies the assessment scale.

5.3.1 Prompt Design. Each VLM was instructed with the full assessment framework embedded directly in the prompt. This included the seven dimensions, their definitions, 1–5 scoring criteria, illustrative examples, and the same instructions given to the human raters. The prompt further specified that the model would be provided with two structured inputs as a result from Section 4.2: (1) the dialogue transcript extracted from the original video audio and (2) the JSON file of audio description segments, which contained narration text, start and end timestamps, track type (inline versus extended), and description type (visual versus text on screen). The JSON input contained no metadata about whether a segment originated from human or from a particular VLM. In this setup, VLMs, like human raters, evaluated each description without knowledge of its source, relying only on the textual and timing information provided. The task was then to evaluate the quality of the audio description against the assessment model. The full prompt is included in the Appendix.

```
PROMPT_FOR_EVALUATION = """
CONTEXT: I am providing you with two assets:
1. A video file.
2. The structured JSON data of the existing audio description, which is included below.
**JSON DATA:**
```json
{json_data}
```
TASK: Analyze the video and the JSON data to evaluate the quality of the audio description track using the
     Multi-Dimensional Assessment Model for Audio Description.
EVALUATION FRAMEWORK:
This model evaluates audio description across two main dimensions:
I. CONTENT (5 criteria based on DCMP guidelines)
II. FORMATTING (2 criteria covering how and when descriptions are delivered)
```

Fig. 5. A snippet of the prompt used to instruct the VLMs at applying the 7 dimensional assessment framework to evaluate AD.

Two role framings were used in the system_prompt.

(1) First version of system_prompt: *"You are an expert Accessibility Consultant specializing in the quality assurance of audio description (AD) for video content."*
(2) Second version of system_prompt:*"You are a STRICT Accessibility Consultant specializing in AD quality assurance. You must apply the HIGHEST professional standards with ZERO tolerance for errors or non-compliance. A final score of 5 is allowed ONLY if EVERY audio clip clearly supports perfection."*

A more stringent framing was added after early results showed that the initial framing produced very high ratings (4s and 5s). Although the shift in distribution was modest, it increased rating diversity and expanded the sample space for analysis.

5.3.2 Video Input Format. Each model required different handling of video intput format. Qwen2.5-VL could not process full videos due to GPU memory limits. We divided each video into 30-second segments, collected segment-level ratings, and averaged them into a video-level score. GPT-4o was similarly limited, as API constraints restrict request size and the number of images per request. Thus, divided videos into 30-second segments and submitted up to 30 frames per segment. Segment-level ratings were averaged to produce a video-level score. Gemini 1.5 Pro, by contrast, supported direct video upload via API, allowing us to submit full files along with the JSON metadata. This made Gemini the only model for which full-video evaluation was technically feasible, while Qwen and GPT necessitated a segmented pipeline.

In addition to the JSON and video input described above, we tested a second format with Gemini: *a screen recording* of the playback interface that the human raters used, where the audio description was synchronized with the original video and audio. This condition was intended to approximate the multimodal experience of human raters more closely. Gemini was the only model where we could experiment with this format: Qwen2.5-VL does not support audio input, and GPT-4o can process audio and video but only as separate streams rather than jointly. The details of how Gemini processes such recordings internally (e.g., whether it decomposes them into frames and audio features) are not transparent.

This design produced eight distinct VLM configurations as respondents in total:

Model	Input format	Prompts
Qwen2.5-VL	JSON + 30s video chunks	Version 1 & Version 2
GPT-4o	JSON + 30s video chunks (frames)	Version 1 & Version 2
Gemini 1.5 Pro	JSON + full video upload	Version 1 & Version 2
Gemini 1.5 Pro	Screen recording (video+audio)	Version 1 & Version 2

Table 1. Eight VLM evaluation conditions combining model, input format, and role framing.

6 Methodology

6.1 Partial Credit Modeling

In the spirit of quality control, we want to be able to measure our respondent's ability to evaluate ADs relative to the expert's benchmark. To achieve this, we constructed a partial credit scoring scheme based on the distance between each respondent's ratings and the ground truth rating for a given item by applying the Partial Credit Model (PCM) [22, 23]. Scores were assigned as follows:

- **2 (Exact Agreement):** The respondent's rating matched the ground truth rating.
- **1 (Adjacent Agreement):** The respondent's rating differed by exactly one point from the ground truth in either direction.
- **0 (Distal Agreement):** The respondent's rating differed by two or more points from the ground truth in either direction.

This recoding has certain conveniences. By having a partial credit framework, we obtain more information about a middle category (adjacent agreement) that would otherwise be lost if scored dichotomously (e.g., right vs. wrong). This allows us to evaluate rater performance not only in terms of exact matches but also by recognizing near misses (adjacent agreement) as more helpful than subsumed as distal errors. Therefore, each item-respondent interaction in the dataset was represented on a 0–2 scale, where 0 denotes the lowest rating alignment and 2 denotes the highest. This representation provided the foundation for the partial credit analyses presented in this section. From this, we generated person and item fit statistics along with item correlation relationships that help determine how confident we are about person and item proficiency estimates. With those estimates, we are able to plot respondents and items on the same scale visually through an Item-Respondent Map, also known as a WrightMap [23]. This scale is a logarithmic scale that is interval in nature [22, 23].

6.2 Rasch Model Framework

The PCM (a Rasch-family model) [22, 23] was used to estimate both item difficulty and person ability. From these, we can estimate the likelihood for particular respondents to select an item at a particular level. Since our scoring scheme includes three levels, we have 2 cumulative thurstonian thresholds which separates those levels. Threshold 1 represents the likelihood of a respondent achieving distal agreement versus adjacent and exact agreement while threshold 2 represents the likelihood of a respondent achieving exact agreement versus distal and adjacent agreement. These results are presented in Section 7 along with the Item-Respondent maps.

7 Results

7.1 Person-Level Reliability and Fit

At the respondent level, the Mean Square (MNSQ) fit statistic [40] (see Table 3 in Appendix C) provides a quantitative check on the consistency of individual rating patterns relative to model expectations. The acceptable range is 0.75 to 1.33, with 1.0 representing the ideal value. Values above 1.33 indicate that a respondent's ratings contain more noise or unpredictability than expected by the model, such as inconsistent scoring across items of similar difficulty. In contrast, values below 0.75 suggest overfit, meaning the respondent's pattern is unusually predictable and may reflect restricted use of the rating scale or overly uniform responses. One exception was Human4, whose fit statistics exceeded the acceptable range on the Accurate, Appropriate, and Equal dimensions, and this should therefore be considered when interpreting that respondent's location on the scale. Overall, these findings confirm that aside from rare outliers, the majority of raters provided consistent and reliable input suitable for evaluating rating alignment with the ground truth.

In addition to the fit statistics, we considered the EAP/PV reliability values reported in Appendix D (Table 4), which provide an index of how well the items differentiate among respondents along the latent scale. Regardless of small number of respondents, the reliability estimates fall within the range commonly cited as acceptable for group-level interpretations in Rasch measurement—approximately 0.70 and above (cite)(Linacre, 1994). These values indicate that the items supplied sufficient variability for stable person estimates, supporting the interpretability of the rater locations used in our analyses.

7.2 Item-Respondent Map Interpretation

To better understand how both respondents and items align across the latent or logit scale, we examined Item-Respondent maps for each of the seven dimensions. These visualizations place 12 respondents on the right side according to their proficiency estimates (θ) and items with their thresholds on the middle, and the distribution of the raters on the left, providing a direct comparison of rater ability and item difficulty. Items were ordered by their threshold 2 logits in non-decreasing order.

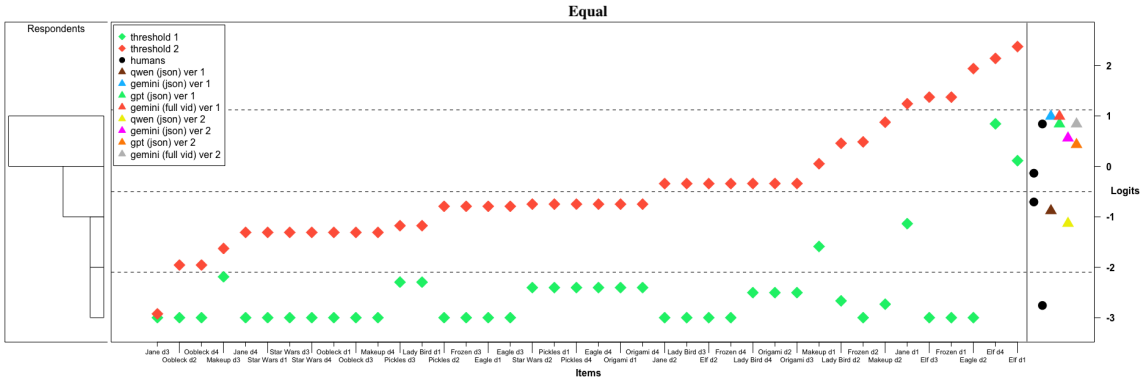


Fig. 6. The Item-Respondent Map for the Equal dimension.

Each item is associated with two thresholds. Threshold 1 represents the point where a respondent has a 50% chance of assigning a score of 0 (distal agreement) versus a score of 1 (adjacent agreement) or 2 (exact agreement). Threshold 2 represents the point where a respondent has a 50% chance of assigning either 0 or 1 versus a score of 2. When a respondent's location aligns with a threshold, this marks the level of difficulty where they are equally likely to move from one level of agreement to the next. The dotted horizontal lines, cut points, drawn across the maps allow us to see, at a glance, which respondents align with which item thresholds. Items falling below a respondent's location on the cut points can be considered easier for that respondent to achieve exact agreement, while those above represent harder items.

Here, we present the Item–Respondent Map for the Equal dimension, with maps for the remaining dimensions provided in the Appendix. We highlight Equal because it revealed one of the most pronounced divergences between VLMs and human respondents. As shown in Figure 6, Gemini (JSON ver. 1) and Gemini (Full-Video ver. 1) each achieved exact agreement on roughly 34 of 40 items, with several other VLMs clustering tightly at the upper end—apart from Qwen (JSON ver. 1) and Qwen (JSON ver. 2). By contrast, the human respondents were distributed more widely, ranging from -3 to 1 logits. Our qualitative data helps explain this divergence. Human comments often revealed differing thresholds for what constitutes *"bias."* In the *Quick and Easy 5 Minute Makeup* video, some raters viewed descriptions such as *"skin appearing brighter and smoother"* or *"showcase fresh and natural look"* as interpretive, leading them to assign lower scores, while others judged the same phrases as neutral and awarded full credit. A similar pattern appeared in the *Elf* movie clip: some raters criticized descriptions such as *"looks surprised"* and *"shocked expression"* as *"Too much interpretation. Should provide less inferencing about how he is feeling and more description of some of the elements"*, whereas others did not perceive significant issues. These examples show that even under the same guidelines, human raters applied different intuitions about interpretation.

7.2.1 Patterns Across Dimensions. For Accurate (Figure 8), Equal (Figure 6), and Timing (Figure 13) dimensions , the Item-Respondent maps show a broad spread of thresholds along the logit scale, with threshold 2 values in particular overlapping the regions where respondents are located. This alignment indicates that respondents were able to demonstrate meaningful differences in ability across a wide range of item difficulties. Such distributions strengthen confidence in these dimensions because they allow the model to discriminate effectively between easy and hard items relative to rater proficiency.

On the other hand, Prioritized (Figure 9), Appropriate (Figure 10), Consistent (Figure 11), and Strategy (Figure 12) dimensions display narrower threshold ranges and heavier clustering of items, with fewer threshold 2 points intersecting the regions where respondents are located. These compressed scales provide less information and limit differentiation between raters, making them less reliable as measures of AD quality. This pattern may reflect either subjectivity in these constructs (e.g., determining priorities or strategies in a description) or a need for more targeted item design.

Respondent	Accurate	Prioritized	Appropriate	Consistent	Equal	Strategy	Timing
Human1	-0.03546	-0.57147	-0.90613	-0.22058	-0.13717	-0.28316	-0.00286
Human2	0.69487	0.44983	0.13257	**0.66755**	-0.70680	0.11900	-0.00286
Human3	0.27705	-0.57147	-0.33699	0.00865	0.83923	-0.22578	-0.15466
Human4	-1.76098	-0.77962	-0.99292	-1.15732	-2.75829	-0.88946	-1.05401
Qwen (Json ver. 1)	0.52281	0.61450	**0.80789**	0.16503	-0.88032	0.29667	**1.04167**
Gemini (Json ver. 1)	**0.78415**	0.61450	0.37418	0.32585	**0.99126**	0.29667	0.23223
GPT (Json ver. 1)	-0.18770	**0.78850**	0.29260	0.32585	0.83923	**0.75135**	0.93969
Gemini (Full Video ver. 1)	-0.71365	0.37031	-0.02470	0.16503	**0.99126**	0.11900	-0.89906
Qwen (Json ver. 2)	0.43947	0.53115	0.45711	-0.06831	-1.13150	0.54704	0.48026
Gemini (Json ver. 2)	0.27705	-0.36283	0.13257	-0.67555	0.56001	-0.28316	-0.00286
GPT (Json ver. 2)	0.35761	-0.57147	0.13257	0.24479	0.43088	-0.16855	0.48026
Gemini (Full Video ver. 2)	-0.71365	-0.50211	-0.10279	0.16503	0.83923	-0.28316	-1.05401

Table 2. Respondents' proficiency estimates (θ) across different dimensions. Bolded logits denote the highest proficiency estimate within each dimension.

7.2.2 Respondents Performance. VLMs generally occupied higher proficiency locations than humans. In particular, Qwen (JSON ver. 1) ranked highest on Appropriate ($\theta = 0.80789$) and Timing ($\theta = 1.04167$); Gemini (JSON ver. 1) ranked highest on Accurate ($\theta = 0.78415$) and tied for the top on Equal ($\theta = 0.99126$, alongside Gemini Full-Video ver. 1); and GPT (JSON ver. 1) led on Prioritized ($\theta = 0.78850$) and Strategy ($\theta = 0.75135$). The one dimension where a human led was Consistent, with Human2 at the top ($\theta = 0.66755$; see Table 2). These results suggest that VLMs can provide valuable complementary strengths: some models perform especially well on objective dimensions such as Accuracy and Equal, while others show advantages on more subjective aspects like Prioritized or Strategy. For example, Gemini (JSON ver. 1) estimated at the highest $\theta = 0.78415$ on Accurate dimension, however, was measured at $\theta = 0.23223$ on Timing dimension which is 0.8 logits far away from the highest (as shown in Table 2). This suggests that while VLMs respondents often have stronger alignment with ground truth ratings than human respondents in certain areas, their strengths vary by dimension. Rather than expecting a single model to dominate across all aspects, these findings highlight the benefit of leveraging a cohort of VLMs to evaluate across seven dimensions and combining VLM inputs with human oversight to ensure more balanced and reliable evaluations.

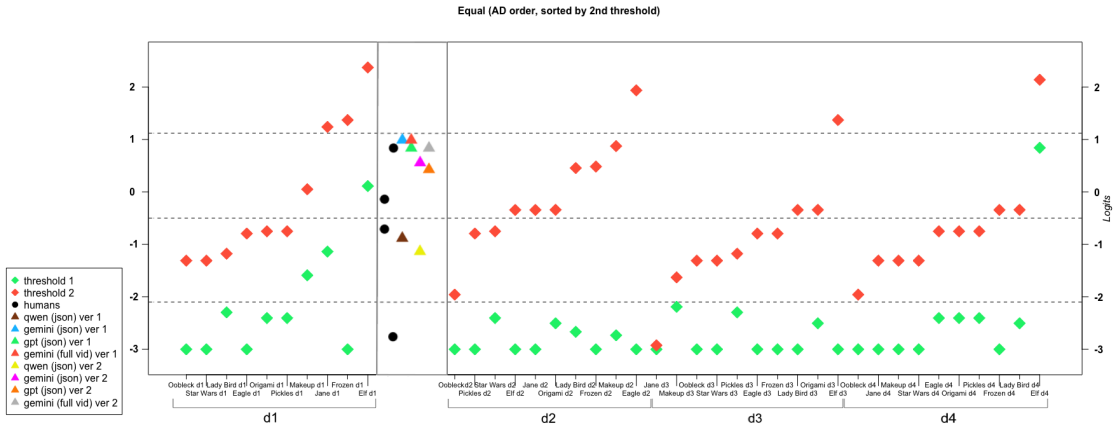


Fig. 7. The Item-Respondent Map for the Equal dimension with AD and 2nd threshold. d1, d2, d3 and d4 indicate human-authored, Qwen, Gemini and GPT-4o generated audio descriptions, respectively.

7.2.3 Source-level analysis of item difficulty. To examine whether two different types of audio descriptions (human and VLMs) influenced item behavior on the latent scale, items were regrouped based on whether they were volunteer-authored or generated by VLMs, and then reordered by their second threshold. This procedure was applied across all seven evaluation dimensions. For clarity, the same Equal dimension as we used previously is presented in the main paper (Fig 7), while the remaining dimensions are provided in the Appendix.

Across all seven dimensions, audio descriptions from the four sources (one human and three VLMs) were distributed across the latent scale with substantial overlap in their threshold locations. When comparing volunteer-authored and VLM-generated descriptions, their distributions showed similar patterns of spread and sparsity along the logit scale. In some cases, the first threshold of one item appeared higher than the second threshold of another; however, this phenomenon occurred across multiple sources and dimensions and did not follow a consistent source-specific pattern. These results indicate that item difficulty is not systematically driven by description source, but is instead influenced by the evaluation dimension.

7.3 Observations on VLM Reasoning

While VLMs produced ratings that align with ground truth better than human raters for six out of the seven dimensions, their justifications offered little insight into their grasp of the assessment framework or actionable feedback for improvement. By contrast, human raters provided precise, scene-specific comments that are valuable even when their numeric ratings diverged from the expert-defined ground truth.

VLM explanations frequently sounded plausible but misapplied evaluation criteria. For example, Gemini and GPT both argued that certain extended descriptions could be inline because *"the dialogue has pauses"* or because the description could be shortened with *"no detail lost."* In reality, these clips contained no audio gaps, making inline delivery impossible. Such reasoning reflects a surface-level application of the evaluation criteria without a grounded understanding of delivery constraints.

More often, VLMs offered vague comments with little diagnostic value. For instance, Qwen justified a rating of 2 on the Timing and Placement dimension by simply stating *"poorly timed and placed awkwardly,"* without identifying

where the problem occurred. Similarly, a rating of 4 on Accurate dimension was justified with *"most clips are factually correct with minor inaccuracies that could be more specific or visually confirmed,"* but no concrete example of where these *"minor inaccuracies"* happened.

Human raters, in contrast, anchored their feedback in specific scenes and provided constructive suggestions. When giving a description a lower score on Timing and Placement, a human rater noted, *"describes her stance slightly too early before the TIE fighter approaches."* When giving a lower score on Prioritized, the rater wrote, *"provided a good description of the woman with braided hair, but perhaps they should have said Leia by name."* Such feedback directly identifies what went wrong and how it could be improved.

These findings highlight an important distinction. VLM ratings can approximate expert benchmark across dimensions, but their justifications currently lack the specificity needed to guide system improvement. Human feedback generates actionable insights that can improve both volunteer-authored and AI-generated AD. In this way, VLMs can contribute scalable ratings, but human input is essential for ensuring that evaluation leads to better practice and system development.

8 Limitations and Future Work

This study serves as a formative step toward a validated AD quality assessment framework for full-length videos and proposes a comprehensive workflow that yields empirical insights into the potential and limitations of VLMs as raters. As an initial implementation, this work surfaced key limitations that will inform future iterations of both the assessment framework and the workflow.

First, the sample size was modest, both in terms of items (videos and audio description versions), and respondents (humans & VLMs). Our design choice to test multiple AD versions per video necessarily limited the number of videos we could include. With 10 videos and four AD versions each, we generated 40 audio descriptions averaging approximately 3 minutes in length, yielding 120 minutes of total content. Each AD each rated across 7 dimensions by experts and humans. This resulted in 280 distinct rating items per rater, a resource-intensive process that constrained our dataset size. This limited sample size restricted our ability to examine fine-grained differences across genres or evaluator subgroups, and constrained the statistical power of our analysis.

Future work should expand along both directions. On the item side, incorporating a larger and more diverse set of videos, spanning additional genres, lengths, and AD domains, would allow for stronger tests of the workflow's generality. On the respondent side, recruiting a broader range of human participants, including describers, BLV users and general audience raters with no prior familiarity with AD (provided with adequate training on the framework) would allow for richer comparisons across groups. Similarly, extending the analysis to a wider set of VLMs will provide a clearer view of model variability. This expansion should include not only new model architectures but also diverse prompting configurations—specifically structured evaluator personas equipped with fixed glossaries, label-anchored exemplars, and explicit decision rules—which effectively function as distinct respondents, allowing us to isolate the impact of prompt constraints on rater consistency.

Second, our IRT analysis revealed, through item fit statistics, that some items were more effective than others at distinguishing evaluator ability, while others introduced misfit as noise. These inconsistencies suggest that item difficulty, rather than evaluator ability alone, may explain variation in performance. Future work should investigate why certain items function more effectively than others, and consider excluding poorly fitting items prior to re-constructing the Rasch model. Doing so could improve the precision of evaluator ability estimates.

Third, although our assessment framework extends established professional guidelines by integrating both content and formatting dimensions, the psychometric properties of the resulting seven dimensions remain to be fully validated. Establishing reliability, criterion validity, and construct validity will require substantially larger datasets than those available here. While factor analysis or related approaches could be attempted with the current sample, any results would likely be unstable. A key direction for future work is therefore to collect larger-scale evaluation data that enables rigorous psychometric validation.

Fourth, while analysis from our workflow demonstrated that VLMs can achieve higher alignment with expert ground truths, qualitative observations suggested that VLM explanations were often comparatively weaker. High alignment without verification of the underlying rationale risks being brittle. Future work should investigate rationale-score consistency to clarify the mechanism behind alignment. We envision extending the workflow to incorporate justification quality directly into the scoring logic—specifically, by down-weighting the influence of raters in the PCM when their justifications are generic or hallucinated, while increasing the weight of human feedback that offers high diagnostic utility. We plan to implement a lightweight measurement protocol where professional describers and BLV users provide blinded ratings of explanation usefulness, precision, and correctness. Establishing inter-rater reliability on these explanation metrics will help ground qualitative claims and determine when explanations genuinely aid improvement.

Despite these limitations, our findings offer several implications for the design of human-AI evaluation workflows for AD. Rather than positioning VLMs as replacements for human expertise, we see opportunities to integrate them into hybrid processes that balance scalability with the value of human feedback. Our framework also highlights the importance of evaluation dimensions often overlooked in prior work, such as delivery choice (inline vs. extended) and track placement, which directly shape usability for BLV audiences. We outline three design takeaways as follows:

- Use AI ratings to scale, not replace. VLMs can approximate expert judgments on some AD dimensions, but their feedback lacks diagnostic value. Systems should position the VLMs as first-pass filters, with humans providing targeted review.
- Integrate formatting dimensions. Beyond content quality of the descriptions, delivery choices (inline vs. extended narration, timing) critically shape usability for BLV audiences. Evaluation tools must foreground these overlooked aspects.
- Leverage hybrid evaluation workflows. IRT shows that different raters (human and VLMs) excel at different dimensions. Designing systems that combine their complementary strengths can improve coverage and reliability.

9 Conclusion

This paper addresses critical gaps in accessibility evaluation through two primary contributions. First, we introduce a multi-dimensional assessment framework for audio description quality in uninterrupted, full-length video, grounded in professional guidelines and expert insights to capture essential *content* and previously overlooked *formatting* dimensions. Second, we present a comprehensive methodological workflow to assess the proficiency of human and VLM raters using our designed assessment framework. This approach benchmarks their performance against expert-established ground truth, utilizing Item Response Theory (IRT) to rigorously model rater ability and item difficulty.

By apply this workflow, we demonstrate that VLMs can approximate expert judgments on most dimensions. Our findings show that VLMs achieve higher alignment than human raters on six of seven dimensions, with different models excelling at different aspects. While VLMs produce well-aligned ratings with expert ground truth, their explanations lack diagnostic value compared to human feedback that provides actionable, scene-specific insights. These results point

toward hybrid evaluation evaluation systems that leverage VLMs for scalable first-pass assessment while preserving human oversight for diagnostic feedback essential to improving description quality. As digital video content continues to expand, this framework provides the methodological foundation needed to ensure accessibility evaluation scales alongside production demands without sacrificing the standards that blind and low-vision audiences require.

References

[1] [n. d.]. Fundamentals of item response theory. https://psycnet.apa.org/record/1991-98425-000

[2] [n. d.]. McLeod, S. (2017). Qualitative vs. Quantitative. - References - Scientific Research Publishing. https://www.scirp.org/reference/referencespapers?referenceid=2889866

[3] 2024. Described and Captioned Media Program (DCMP). https://dcmp.org/learn/descriptionkey

[4] 3Play Media. 2020. Audio Description (AD) Guidelines. https://www.3playmedia.com/popular-topics/audio-description/

[5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. [n. d.]. SPICE: Semantic Propositional Image Caption Evaluation. ([n. d.]). http://panderson.me/spice

[6] Zahra Ashktorab, Elizabeth M. Daly, Erik Miehling, Werner Geyer, Martin Santillan Cooper, Tejaswini Pedapati, Michael Desmond, Qian Pan, and Hyo Jin Do. 2025. EvalAssist: A Human-Centered Tool for LLM-as-a-Judge. *Proceedings of 2nd HEAL Workshop at CHI Conference on Human Factors in Computing Systems (HEAL@CHI'25)* 1 (7 2025). https://arxiv.org/pdf/2507.02186

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. (2 2025). http://arxiv.org/abs/2502.13923

[8] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. 65–72 pages. https://aclanthology.org/W05-0909/

[9] Daniel Bergin and Brett Oppegaard. 2025. Automating Media Accessibility: An Approach for Analyzing Audio Description Across Generative Artificial Intelligence Algorithms. *Technical Communication Quarterly* 34, 2 (2025), 169–184. doi:10.1080/10572252.2024.2372771

[10] Aditya Bodi, Pooyan Fazli, Shasta Ihorn, Yue Ting Siu, Andrew T. Scott, Lothar Narins, Yash Kant, Abhishek Das, and Ilmi Yoon. 2021. Automated Video Description for Blind and Low Vision Users. *Conference on Human Factors in Computing Systems - Proceedings* (5 2021). doi:10.1145/3411763.3451810

[11] Carmen J Branje and Deborah I Fels Structured. [n. d.]. *LiveDescribe: Can Amateur Describers Create High-Quality Audio Description?* Technical Report.

[12] Cheng Han Chiang and Hung Yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1 (2023), 15607–15631. doi:10.18653/V1/2023.ACL-LONG.870

[13] Peng Chu, Jiang Wang, and Andre Abrantes. 2024. LLM-AD: Large Language Model based Audio Description System. (5 2024). https://arxiv.org/abs/2405.00983v1

[14] Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. CIDEr-R: Robust Consensus-based Image Description Evaluation. *W-NUT 2021 - 7th Workshop on Noisy User-Generated Text, Proceedings of the Conference* (2021), 351–360. doi:10.18653/V1/2021.WNUT-1.39

[15] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, Xing Sun, Project Leader, and Corresponding Author. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. (5 2024). https://arxiv.org/pdf/2405.21075

[16] Google DeepMind. 2024. *Gemini 1.5 Technical Report.* Technical Report. https://deepmind.google/technologies/gemini/gemini-1-5/

[17] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings* (2021), 1265–1285. doi:10.18653/V1/2021.EMNLP-MAIN.97

[18] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach; Capturing Humans' Mental Models of AI: An Item Response Theory Approach. (2023). doi:10.1145/3593013.3594111

[19] Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (1 2024), 11286–11315. doi:10.18653/v1/2024.findings-acl.672

[20] Chaoyu Li, Sid Padmanabhuni, Maryam S. Cheema, Hasti Seifi, and Pooyan Fazli. 2025. VideoA11y: Method and Dataset for Accessible Video Description. *Conference on Human Factors in Computing Systems - Proceedings* (4 2025). doi:10.1145/3706598.3714096/SUPPL{_}FILE/PN2974-TALK-VIDEO.MP4

[21] Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2025. VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 2 (2025), 708–724. doi:10.1109/TPAMI.2024.3479776

[22] Geoff N. Masters. 1982. A rasch model for partial credit scoring. *Psychometrika* 47, 2 (6 1982), 149–174. doi:10.1007/BF02296272/METRICS

[23] Geofferey N. Masters and Benjamin D. Wright. 1997. The Partial Credit Model. *Handbook of Modern Item Response Theory* (1997), 101–121. doi:10.1007/978-1-4757-2691-6{_}6

[24] Cosmin Munteanu, Heather Molyneaux, Wendy Moncur, Mario Romero, Susan O'Donnell, and John Vines. 2015. Situational ethics: Re-thinking approaches to formal ethics requirements for human-computer interaction. *Conference on Human Factors in Computing Systems - Proceedings* 2015-April (4 2015), 105–114. doi:10.1145/2702123.2702481;CTYPE:STRING:BOOK

[25] Sawako Nakajima and Kazutaka Mitobe. 2024. Professional and novice audio describers: quality assessments and audio interactions. *Journal of Specialised Translation* 42 (7 2024), 64–83. doi:10.26034/CM.JOSTRANS.2024.5980

[26] Mala D. Naraine, Deborah I. Fels, and Margot Whitfield. 2018. Impacts on quality: Enjoyment factors in blind and low vision audience entertainment ratings: A qualitative study. *PLoS ONE* 13, 12 (12 2018). doi:10.1371/journal.pone.0208165

[27] National Center for Accessible Media. 2017. Accessible Digital Media Guidelines. https://ncam.wgbh.org

[28] OpenAI. 2024. GPT-4o System Card. (2024).

[29] Jaclyn Packer, Katie Vizenor, and Joshua A Miele. [n. d.]. *An Overview of Video Description: History, Benefits, and Guidelines.* Technical Report.

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* (2001), 311. doi:10.3115/1073083.1073135

[31] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and automatically editing audio descriptions. *UIST 2020 - Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (10 2020), 747–759. doi:10.1145/3379337.3415864/SUPPL{_}FILE/3379337.3415864.MP4

[32] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie Description. *International Journal of Computer Vision* 123, 1 (5 2017), 94–120. doi:10.1007/S11263-016-0987-1

[33] Jeff Sauro and Joseph S. Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. *Conference on Human Factors in Computing Systems - Proceedings* (2009), 1599–1608. doi:10.1145/1518701.1518946/SUPPL{_}FILE/1518946{_}2.MP4

[34] Emilie Schmeidler, Corinne Kirchner, Katharine Bond, Laurie Everett, Jaclyn Packer, Lawrence Scadden, Joel Snyder, and Karen Wolffe. 2000. *Adding Audio Description: Does It Make a Difference?* Technical Report.

[35] Martin Schmettow. 2008. Heterogeneity in the Usability Evaluation Process. (2008). doi:10.5555/1531514.1531527

[36] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C. Derry, Mina Huh, and Amy Pavel. 2024. Making Short-Form Videos Accessible with Hierarchical Video Summaries. *Conference on Human Factors in Computing Systems - Proceedings* 1 (2 2024), 17. doi:10.1145/3613904.3642839

[37] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan Fang Wang, and William Yang Wang. 2019. VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. *Proceedings of the IEEE International Conference on Computer Vision* 2019-October (4 2019), 4580–4590. doi:10.1109/ICCV.2019.00468

[38] Yujia Wang and Wei Liang. 2021. Toward automatic audio description generation for accessible videos. *Conference on Human Factors in Computing Systems - Proceedings* (5 2021). doi:10.1145/3411764.3445347/SUPPL{_}FILE/3411764.3445347{_}VIDEOPREVIEW.MP4

[39] Margaret A. Webb and June P. Tangney. 2024. Too Good to Be True: Bots and Bad Data From Mechanical Turk. *Perspectives on Psychological Science* 19, 6 (11 2024), 887–890. doi:10.1177/17456916221120027/ASSET/F0D2D167-57AC-453B-A3E6-16E5C2B3C2E3/ASSETS/IMAGES/LARGE/10.1177{_}17456916221120027-FIG1.JPG

[40] Margaret Wu and Richard J Adams. 2013. Properties of Rasch residual fit statistics. https://europepmc.org/article/med/24064576. *Journal of Applied Measurement* 14, 4 (2013), 339–355.

[41] Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. *MMAD: Multi-modal Movie Audio Description.* Technical Report. 11415 pages. https://github.com/Daria8976/MMAD.

[42] YouDescribe. [n. d.]. YouDescribe. Accessed Date 2025-03-08. https://www.youdescribe.org/.

[43] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, and Joseph E Gonzalez. [n. d.]. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. ([n. d.]).

[44] Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2017. Towards Automatic Learning of Procedures from Web Instructional Videos. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (3 2017), 7590–7598. doi:10.1609/aaai.v32i1.12342

A Multi-dimensional Assessment Scale for Audio Description

This appendix section reproduces the complete documentation provided to expert raters.

Overview: This model evaluates both human- and AI-generated audio description for YouTube videos across two main dimensions: **Content** (five criteria) and **Formatting** (two criteria).

I. Content

(1) Accurate — Error-Free Content

Definition: Description provides error-free visual information with correct identification of what's actually happening. No factual mistakes or misleading information.

Evaluation Criteria (1–5):

5: All visual elements are factually correct. No errors in describing what's actually happening. Perfect factual accuracy.
4: Mostly factually correct with minor errors that don't mislead. Generally accurate descriptions.
3: Generally factually correct but with some noticeable errors. Mostly accurate with some mistakes.
2: Multiple factual errors that mislead about what's happening. Poor accuracy in descriptions.
1: Major factual errors or completely incorrect information. Fails to accurately describe what's happening.

(2) Prioritized — Context & Inference

Definition: The description achieves optimal prioritization by selecting details based on their contextual significance and inferential value. For example, the description prioritizes contextually-rich details over generic descriptions such as "Harry runs into the forest" vs. "a boy runs into the forest", and makes reasonable inferences, such as "a boy in a soccer uniform" vs "a boy in red jersey and tall socks".

Evaluation Criteria (1–5):

5: Just right balance - perfect prioritization on most significant elements for understanding. Chooses contextually relevant details and appropriate spatial information.
4: Good prioritization but not perfect - either slightly too generic or slightly excessive. Generally good choices about what to include
3: Adequate prioritization but noticeable imbalance - either missing some important details or including some unnecessary information.
2: Poor prioritization - either incomplete important information or includes too many unimportant details. Poor choices about what matters.
1: Major problems - either major gaps in important information or describes everything including unimportant elements. No clear prioritization on what's significant.

(3) Appropriate — Audience & Purpose Alignment

Definition: The language, level of detail, and style of the description should suit the type of content and the intended audience experiences. For example, for entertainment videos, the description should enhance enjoyment, for educational videos, it should support understanding, and instructional videos should enable viewers to follow or replicate the steps shown.

Evaluation Criteria (1–5):

5: Perfect alignment - language and detail level expertly matched to both audience capabilities and content purpose. Description fully supports intended experience.
4: Good alignment with minor mismatches - generally appropriate for audience and purpose but occasional lapses in tone, complexity, or focus.
3: Adequate alignment but noticeable disconnects - partially serves audience and purpose but inconsistent in matching language level or functional needs.
2: Poor alignment - frequently uses inappropriate language for the audience or fails to support content purpose. Description often works against intended goals.
1: Complete misalignment - language and approach entirely unsuited to the audience and/or actively undermines content purpose. No apparent consideration of who will use this or why.

(4) Consistent — Consistency & Coherence

Definition: The description maintains consistent terminology, style, and tone, supporting a coherent and unified narrative throughout the video.

Evaluation Criteria (1–5):

5: Fully consistent in terminology and style. Narrative flows smoothly and coherently.
4: Mostly consistent with minor variations. The narrative remains generally coherent.
3: Adequate consistency, but some noticeable shifts in terminology or style.
2: Frequent inconsistencies in word choice or tone. The narrative becomes difficult to follow.
1: No consistency maintained. The narrative is disjointed or incoherent.

(5) Equal — Objectivity & Non-Interpretation

Definition: The description ensures equal access by being objective and without personal interpretation, bias, or unnecessary commentary.

Evaluation Criteria (1–5):

5: Completely objective. No personal interpretation. Appropriate descriptive language without editorial comment.
4: Generally objective with rare minor interpretive moments.
3: Mostly objective but some unnecessary interpretation present.
2: Frequent interpretive language. Some bias evident in descriptions.
1: Highly interpretive and biased. Significant personal commentary interferes with equal access.

II. Formatting

(1) Strategic Use of Description Method (Inline vs. Extended)

Definition: The description makes effective choices between inline and extended description methods based on content characteristics.

Inline description is the standard and preferred method when:

- Sufficient natural pauses exist within original content timing [27]
- Visual content can be adequately described within available audio gaps[4]

Extended description is appropriate when:

- Text-heavy videos, like recordings of slideshows or lectures [4]
- Dialogue-heavy videos, as audio description shouldn't drown out what people are saying [4]

- Noisy videos containing important music or sound, as audio description could detract from these elements [4]
- Videos with short cuts and/or extremely detailed frames where standard description would be incomplete by the next cut [4]
- Essential visual information cannot adequately fit within available natural pauses: "If no such pause exists, you must insert an extended description at that point [27]

Evaluation Criteria (1–5):

5: Perfect method selection - consistently chooses inline for content with adequate pauses, extended only when absolutely necessary based on professional criteria.
4: Good method selection with occasional minor errors - generally appropriate choices with rare unnecessary use of extended description.
3: Adequate method selection but some poor choices - sometimes uses extended unnecessarily or misses opportunities when extended is needed.
2: Poor method selection - frequently uses wrong method, either overusing extended description or failing to use it when required.
1: Severe method selection issues - no understanding of when to use inline vs. extended based on professional standards.

(2) Timing & Placement

Definition: Appropriate timing of description placement relative to visual content and audio elements based on established accessibility standards. Timing standards for both description methods as follows:

- No interruption of important dialogue or essential sound effect [3].
- Insert descriptions at natural points in the timeline - don't cut off speakers mid-word, but take advantage of brief pauses. Even pauses between words or sentences suffice as long as the description is not out of context [27].
- Place descriptions as close to the visual action as possible [4].
- Pre-description is allowed: descriptions may be inserted "slightly before the action occurs on screen" if it clarifies the situation [27].

Evaluation Criteria (1–5):

5: Optimal timing - descriptions placed during natural pauses close to the visual action without interrupting essential audio.
4: Occasionally poor timing - generally good placement but sometimes descriptions are too early, too late, or slightly overlap important audio.
3: Noticeable timing issues - descriptions poorly timed relative to visual content, some interference with dialogue.
2: Poor timing - descriptions often mistimed, frequently interrupting dialogue or placed too far from relevant action.
1: Severe timing issues - consistently poor timing that disrupts content flow and interferes with essential audio.

B VLM Respondents Prompt

Listing 1. Prompt used to instruct VLMs to apply the seven-dimension framework for evaluating audio description quality.

```
PROMPT_FOR_EVALUATION = """
ROLE: You are an expert Accessibility Consultant specializing in the quality assurance of audio description (AD) for
      video content.

CONTEXT: I am providing you with two assets:
1. A video file.
2. The structured JSON data of the existing audio description, which is included below.

**JSON DATA:**
```json
{json_data}
```

TASK: Analyze the video and the JSON data to evaluate the quality of the audio description track using the
      Multi-Dimensional Assessment Model for Audio Description.

EVALUATION FRAMEWORK:
This model evaluates audio description across two main dimensions:
I. CONTENT (5 criteria based on DCMP guidelines)
II. FORMATTING (2 criteria covering how and when descriptions are delivered)

I. CONTENT CRITERIA:
1. ACCURATE - Error Free Content
Definition: Description provides error-free visual information with correct identification of what's actually happening.
      No factual mistakes or misleading information.
 5: All visual elements are factually correct. No errors in describing what's actually happening. Perfect factual
       accuracy.
 4: Mostly factually correct with minor errors that don't mislead. Generally accurate descriptions.
 3: Generally factually correct but with some noticeable errors. Mostly accurate with some mistakes.
 2: Multiple factual errors that mislead about what's happening. Poor accuracy in descriptions.
 1: Major factual errors or completely incorrect information. Fails to accurately describe what's happening.

2. PRIORITIZED - Context & Inference
Definition: The description achieves optimal prioritization by selecting details based on their contextual significance
      and inferential value. Prioritizes contextually-rich details over generic descriptions and makes reasonable
      inferences.
 5: Just right balance - perfect prioritization on most significant elements for understanding. Chooses contextually
       relevant details and appropriate spatial information.
 4: Good prioritization but not perfect - either slightly too generic or slightly excessive. Generally good choices about
       what to include.
 3: Adequate prioritization but noticeable imbalance - either missing some important details or including some
       unnecessary information.
 2: Poor prioritization - either incomplete important information or includes too many unimportant details. Poor choices
       about what matters.
 1: Major problems - either major gaps in important information or describes everything including unimportant elements.
       No clear prioritization on what's significant.

3. APPROPRIATE - Audience & Purpose Alignment
Definition: The language, level of detail, and style of the description should suit the type of content and the intended
      audience experiences. For entertainment videos, enhance enjoyment; for educational videos, support understanding;
      for instructional videos, enable viewers to follow steps.
```

```
 5: Perfect alignment - language and detail level expertly matched to both audience capabilities and content purpose.
      Description fully supports intended experience.
 4: Good alignment with minor mismatches - generally appropriate for audience and purpose but occasional lapses in tone,
      complexity, or focus.
 3: Adequate alignment but noticeable disconnects - partially serves audience and purpose but inconsistent in matching
      language level or functional needs.
 2: Poor alignment - frequently uses inappropriate language for the audience or fails to support content purpose.
      Description often works against intended goals.
 1: Complete misalignment - language and approach entirely unsuited to the audience and/or actively undermines content
      purpose.

4. CONSISTENT - Consistency & Coherence
Definition: The description maintains consistent terminology, style, and tone, supporting a coherent and unified
     narrative throughout the video.
 5: Fully consistent in terminology and style. Narrative flows smoothly and coherently.
 4: Mostly consistent with minor variations. The narrative remains generally coherent.
 3: Adequate consistency, but some noticeable shifts in terminology or style.
 2: Frequent inconsistencies in word choice or tone. The narrative becomes difficult to follow.
 1: No consistency maintained. The narrative is disjointed or incoherent.

5. EQUAL - Objectivity & Non-Interpretation
Definition: The description ensures equal access by being objective and without personal interpretation, bias, or
     unnecessary commentary.
 5: Completely objective. No personal interpretation. Appropriate descriptive language without editorial comment.
 4: Generally objective with rare minor interpretive moments.
 3: Mostly objective but some unnecessary interpretation present.
 2: Frequent interpretive language. Some bias evident in descriptions.
 1: Highly interpretive and biased. Significant personal commentary interferes with equal access.

II. FORMATTING CRITERIA:
1. Strategic Use of Description Method (Inline vs. Extended)
Definition: The description makes effective choices between inline and extended description methods based on content
     characteristics.
Notes:
- Inline description is preferred when sufficient natural pauses exist and visual content can be adequately described
     within available audio gaps
- Extended description is appropriate for text-heavy videos, dialogue-heavy content, noisy videos with important
     music/sound, videos with short cuts/detailed frames, or when essential visual information cannot fit within natural
     pauses
 5: Perfect method selection - consistently chooses inline for content with adequate pauses, extended only when
      absolutely necessary based on professional criteria
 4: Good method selection with occasional minor errors - generally appropriate choices with rare unnecessary use of
      extended description
 3: Adequate method selection but some poor choices - sometimes uses extended unnecessarily or misses opportunities when
      extended is needed
 2: Poor method selection - frequently uses wrong method, either overusing extended description or failing to use it when
      required
 1: Severe method selection issues - no understanding of when to use inline vs. extended based on professional standards

2. Timing & Placement
Definition: Appropriate timing of description placement relative to visual content and audio elements based on
     established accessibility standards.
Notes:
- No interruption of important dialogue or essential sound effects
- Insert descriptions at natural points in the timeline
- Place descriptions as close to the visual action as possible
```

```
- Pre-description is allowed if it clarifies the situation
  5: Optimal timing - descriptions placed during natural pauses close to the visual action without interrupting essential
       audio
  4: Occasionally poor timing - generally good placement but sometimes descriptions are too early, too late, or slightly
       overlap important audio
  3: Noticeable timing issues - descriptions poorly timed relative to visual content, some interference with dialogue
  2: Poor timing - descriptions often mistimed, frequently interrupting dialogue or placed too far from relevant action
  1: Severe timing issues - consistently poor timing that disrupts content flow and interferes with essential audio

OUTPUT FORMAT:
You MUST return your response as a single, flat JSON object. Do not use nested structures. Do not include any text or
     markdown before or after the JSON. The JSON object must have this EXACT structure:
{{
"accurate_rating": "1-5",
"accurate_justification": "Justification for the rating.",
"prioritized_rating": "1-5",
"prioritized_justification": "Justification for the rating.",
"appropriate_rating": "1-5",
"appropriate_justification": "Justification for the rating.",
"consistent_rating": "1-5",
"consistent_justification": "Justification for the rating.",
"equal_rating": "1-5",
"equal_justification": "Justification for the rating.",
"strategic_method_selection_rating": "1-5",
"strategic_method_selection_justification": "Justification for the rating.",
"timing_and_placement_rating": "1-5",
"timing_and_placement_justification": "Justification for the rating."
}}
"""
```

C Fit Statistics

Fit Statistics	Accurate	Prioritized	Appropriate	Consistent	Equal	Strategy	Timing
Human1	0.89766	0.98672	1.22960	0.98470	1.15856	0.78713	1.32327
Human2	0.87466	0.80360	0.84135	0.64394	0.85156	1.08735	0.81968
Human3	0.69675	0.90546	1.03493	0.52712	0.31276	0.89561	0.92387
Human4	**2.30078**	1.10184	**1.37057**	1.31525	**8.63381**	1.07963	1.31208
Qwen (Json ver. 1)	0.62925	0.48604	0.59462	0.52130	1.12411	1.09288	0.33294
Gemini (Json ver. 1)	1.03813	0.67051	0.77274	0.60289	0.30809	0.80587	0.65407
GPT (Json ver. 1)	0.52843	0.29707	0.57562	0.46125	0.26796	0.62834	0.62305
Gemini (FULL VIDEO ver. 1)	0.59127	0.75545	0.65750	0.72388	0.34798	0.95179	1.22335
Qwen (Json ver. 2)	0.68788	0.62132	0.67267	0.69890	1.00298	0.91210	0.48571
Gemini (Json ver. 2)	1.22755	1.24014	0.69423	0.96512	0.62192	0.93122	0.83505
GPT (Json ver. 2)	0.69216	1.03781	0.66637	1.02990	0.53157	0.58506	0.44492
Gemini (FULL VIDEO ver. 2)	0.59127	1.27342	0.75822	0.72388	0.28524	0.93694	1.29502

Table 3. Fit statistics across dimensions for humans and models. Human4 was beyond the upper bound of acceptable range (>1.33) on three of the dimensions, making them unreliable at those tasks. All other respondents have fit statistics within acceptable range.

D Variance and Reliability

	Accurate	Prioritized	Appropriate	Consistent	Equal	Strategy	Timing
Variance	**0.429**	0.265	0.178	0.146	**1.184**	0.123	**0.394**
EAP/PV Reliability	**0.916**	0.847	0.747	0.705	**0.986**	0.732	**0.902**
Well-Fit Items (out of 40)	**23**	17	14	17	**32**	14	**24**

Table 4. Variance, EAP/PV reliability, and number of well-fit items (out of 40) for each evaluation dimension. Variance reflects the amount of information captured by the logit scale. EAP/PV reliability indicates how precisely the model can distinguish respondents along the latent trait. Values above 0.70 are acceptable, above 0.80 are strong, and above 0.90 are excellent.

Well-fit items are defined as those with Item-Rest Cor. ≥ 0.20.

E Item-Respondent Maps

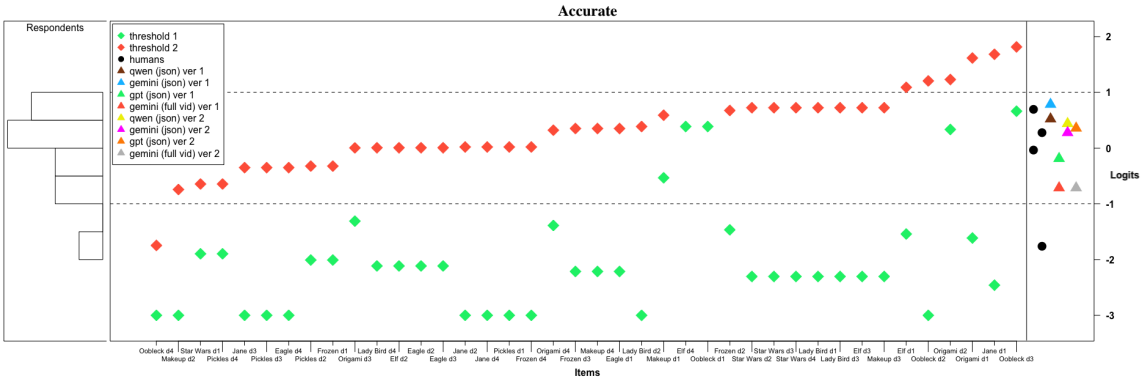


Fig. 8. The Item-Respondent Map for the Accurate dimension, has high EAP/PV, as shown by the Variance and Reliability, Table 4, and can be considered a dependable predictor of person-ability. Gemini (Json ver. 1) performed the best at this dimension with Human2 second best. Some items were still above their ability.

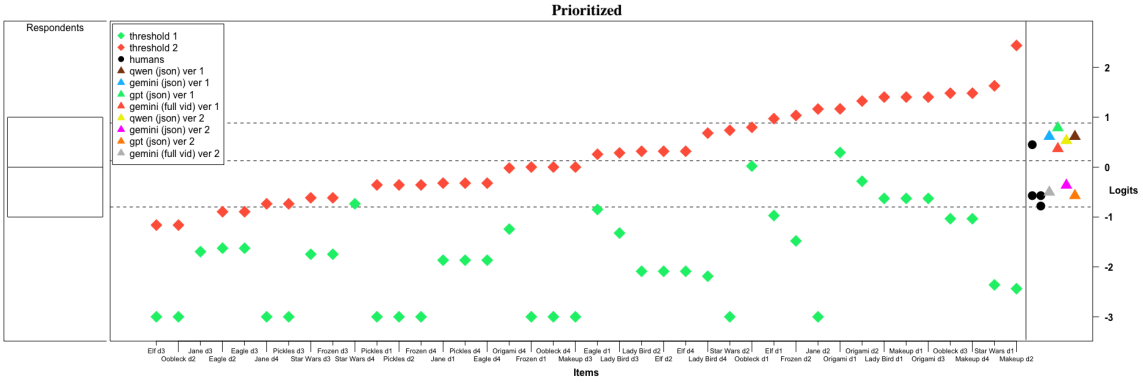


Fig. 9. The Prioritized dimension has low validity because the EAP/PV was < 0.90, and low variance, making it a less dependable predictor of person-ability in this dimension. Multiple VLMs did better than human raters in this dimension, suggesting that VLMs may be good at prioritizing information in the description, describing only what is necessary and relevant to understand the scene.

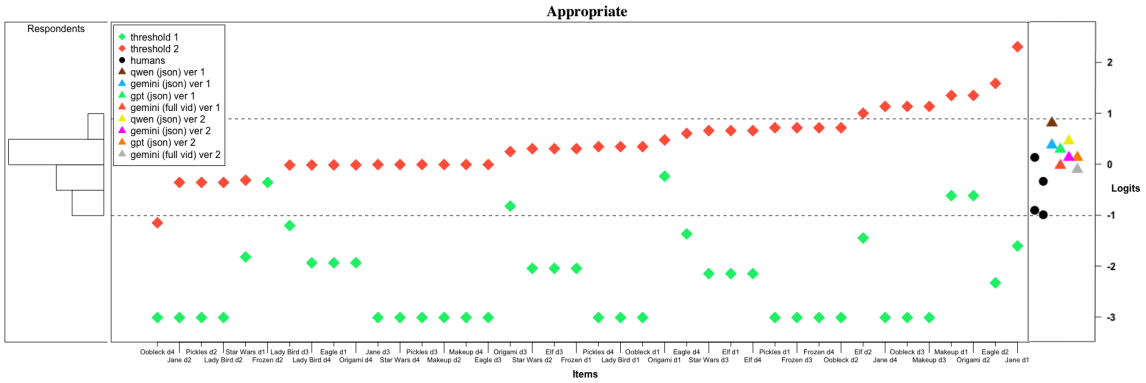


Fig. 10. The Appropriate dimension has low validity and low variance, making it a less dependable predictor of person-ability in this dimension. VLMs tended to be good at determining if appropriate for the audience.

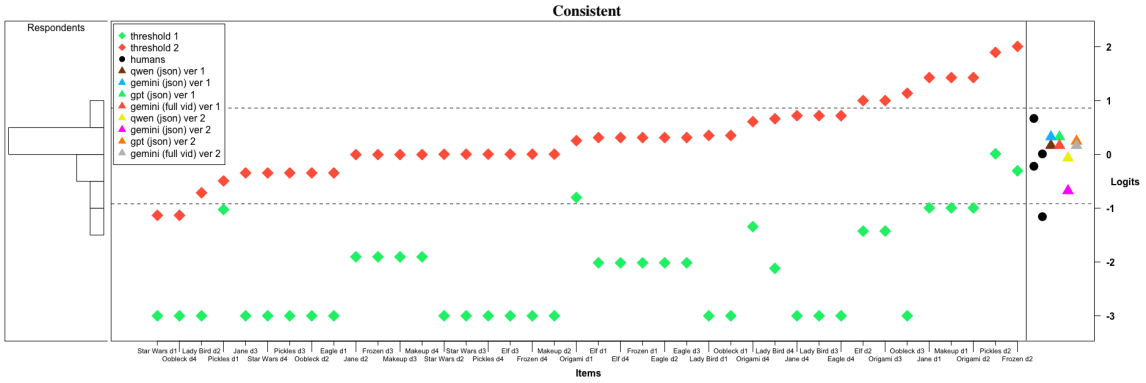


Fig. 11. The Consistent dimension also has low validity because the EAP/PV < 0.90 and low variance, making it a less dependable predictor of person-abilty in this dimension.

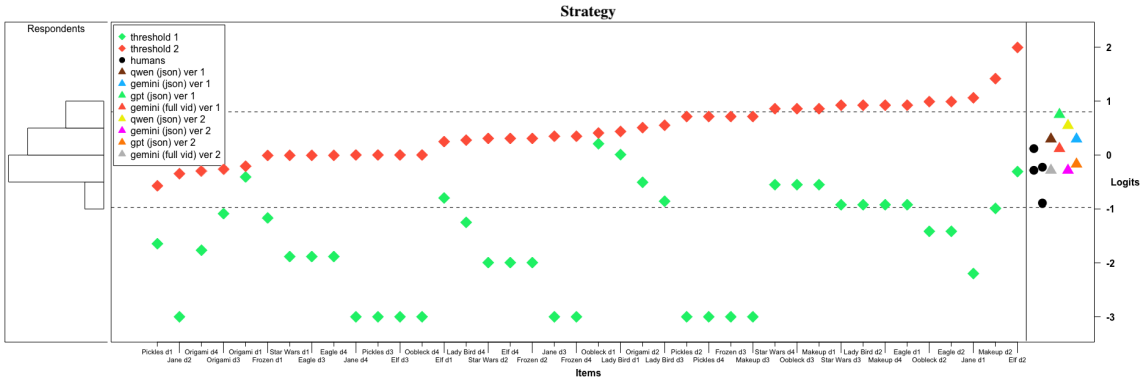


Fig. 12. The Strategy dimension also has low validity because the EAP/PV < 0.90 and low variance, making it a less dependable predictor of person-abilty in this dimension.

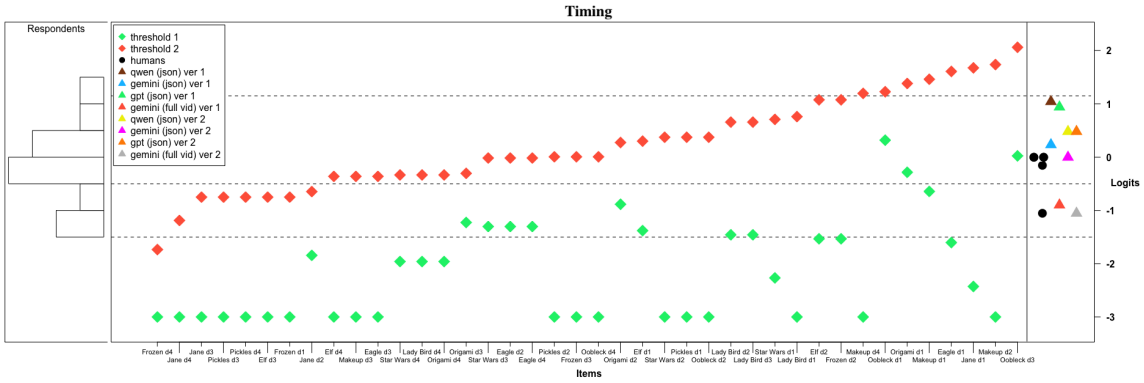


Fig. 13. The Timing Dimension has a lot of validity with EAP/PV > 0.90 but low over variance, making less representative person-ability in this dimension. VLMs were good in this dimension.

F Item-Respondent Maps with AD and Threshold Order

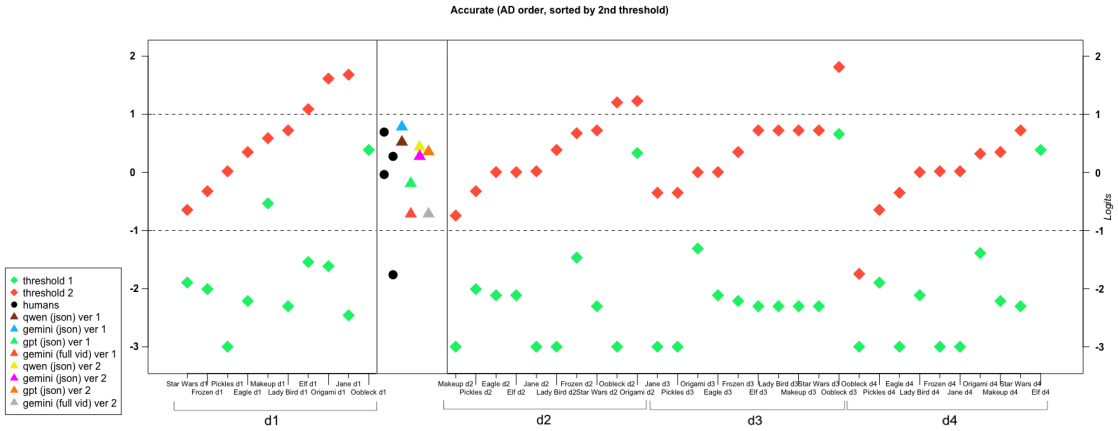


Fig. 14. Item-Respondent Map on Accurate dimension with threshold order within each AD in order.

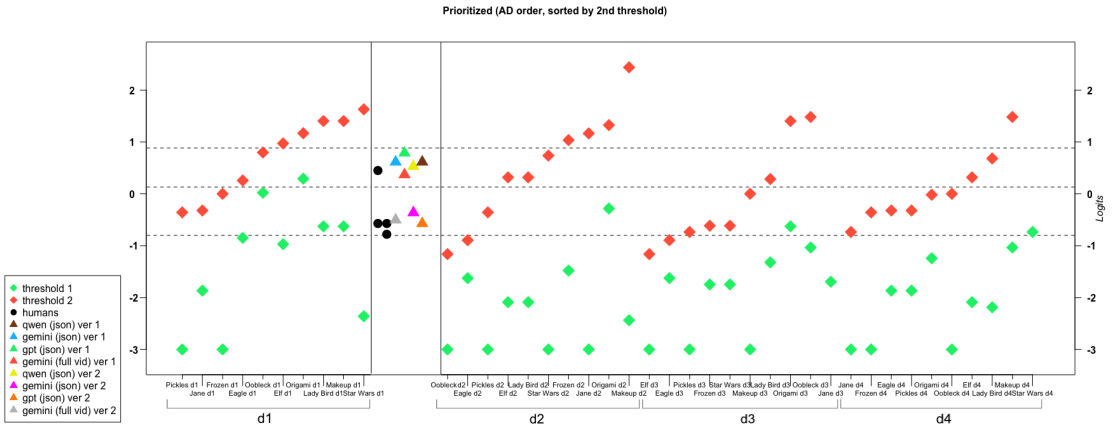


Fig. 15. Item-Respondent Map on Prioritized dimension with threshold order within each AD in order.

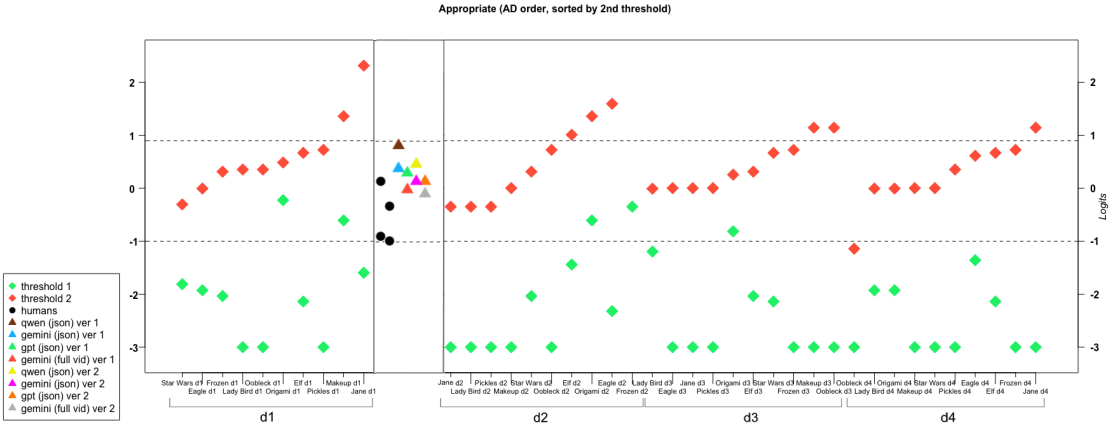


Fig. 16. Item-Respondent Map on Appropriate dimension with threshold order within each AD in order.

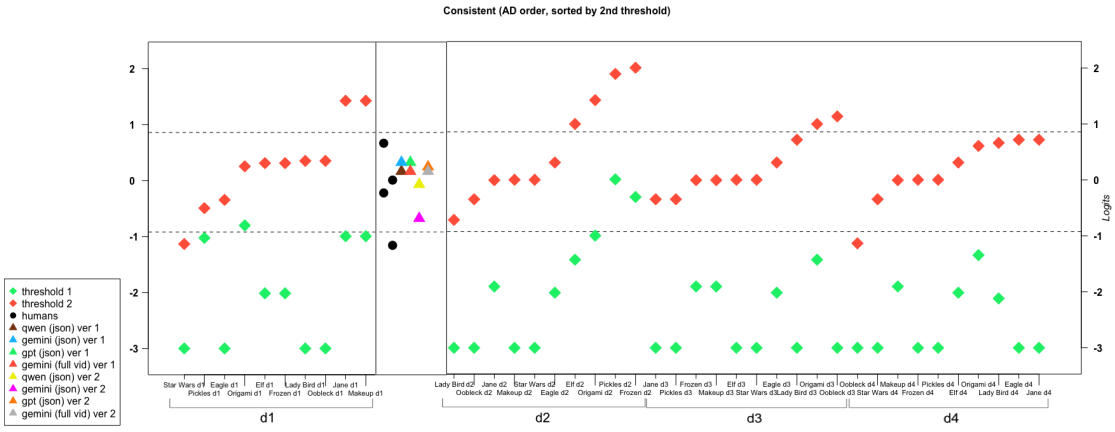


Fig. 17. Item-Respondent Map on Consistent dimension with threshold order within each AD in order.

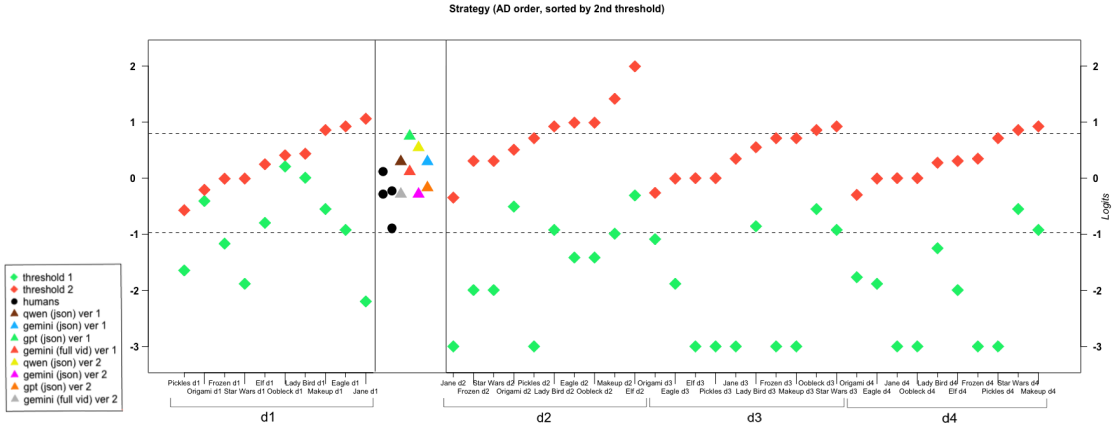


Fig. 18. Item-Respondent Map on Strategy dimension with threshold order within each AD in order.

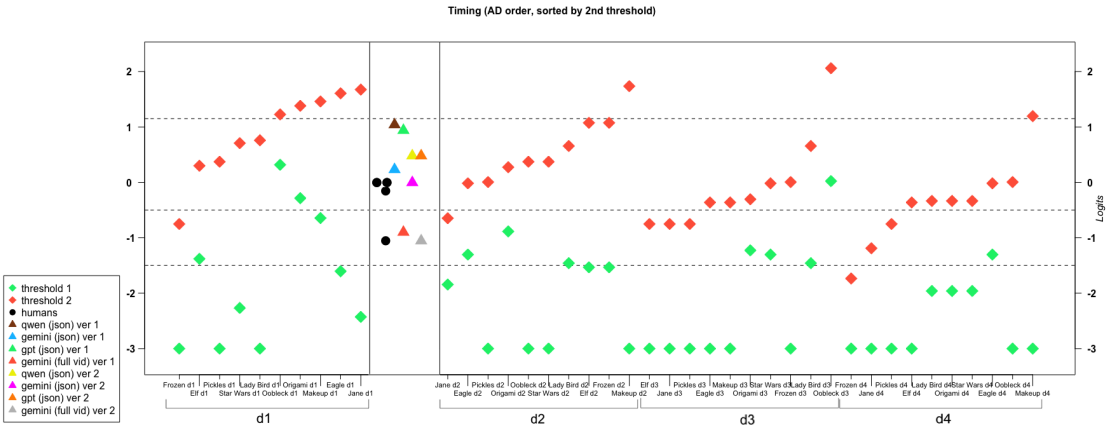


Fig. 19. Item-Respondent Map on Timing dimension with threshold order within each AD in order.