Evaluation of Vision Language Models using Item Response Theory

A Thesis submitted to the faculty of

San Francisco State University

In partial satisfaction of the

requirements for

the Degree

Masters of Science

in

Computer Science

by

Gio Jung

San Francisco, California

December 2025

Copyright by

Gio Jung

2025

Abstract

Evaluating the performance of generative Artificial Intelligence (AI) systems remains a central challenge in computer science. Traditional evaluation metrics such as F1, accuracy, BLEU, METEOR, or CIDEr, treat all items as equally difficult and often fail to capture nuanced differences in model performance. This limitation is especially apparent when assessing Vision Language Models (VLMs), whose outputs must be judged for quality, coherence, and human alignment. Although Item Response Theory (IRT) is widely used in psychometrics to measure human ability and item difficulty, it has not been systematically applied in machine learning evaluation. This thesis advances IRT as both a rigorous approach and a promising direction for evaluating VLMs as raters, offering a measurement-based framework that improves both the validity and interpretability of evaluation results.

Three case studies ground this work. The first examines VLMs rating image captions, comparing model ratings to human judgments using the Validated Image Caption Rating (VICR) dataset. The second extends IRT to visual reading comprehension, where VLMs and human participants interpret comic-based narratives requiring both local and global inference. The third investigates video audio description rating, a complex, multi-dimensional task critical for accessibility, where VLMs are assessed alongside human raters across seven evaluation dimensions. In all cases, Wright Map analyses provide interpretable visualizations that re-

veal meaningful differences between models, show which models approximate human raters most closely, and tells which items are easy or hard to be rated.

This thesis contributes a generalizable framework for evaluating AI models as human-like raters. By applying IRT to VLM evaluation, I demonstrate how psychometric tools can provide valuable insights beyond traditional metrics, enhancing our ability to assess interpretability, reliability, fairness, and alignment. Ultimately, this work positions IRT as both an effective methodology and a forward-looking direction for evaluating VLMs as judges within broader AI workflows.

Acknowledgments

This work would not have been possible without the generous guidance, encouragement, and collaboration of interdisciplinary team from Youdescribe.

First and foremost, I would like to express my deepest gratitude to Dr. Ilmi Yoon, who has been leading this research with vision and passion. Her constant encouragement, clear direction, and trust in my independence allowed me to explore creative ideas and navigate challenges with confidence. This thesis would not have been possible without her mentorship and belief in my potential.

I am also sincerely thankful to Dr. Alexander Blum, whose expertise in measurement and psychometrics was indispensable to this work. His clear explanations and patient guidance helped me develop a deeper understanding of Item Response Theory and its applications.

Special thanks to Andrew Scott for maintaining the dynamic virtual lab environment, which provided invaluable opportunities for technical discussions and collaborations. The open and supportive nature of that community helped me grow both technically and intellectually.

I would also like to thank Dr. Shasta Ihorn for sharing her expertise in statistics, user studies, and inter-rater agreement. Her detailed advice was instrumental in refining the methodological rigor of this research.

My appreciation also goes to Dr. Vassilis Athitsos from the University of Texas at Arlington. His sharp analytical insights and mathematically grounded questions often revealed

aspects I had overlooked and inspired new directions in this project.

I am grateful to my friends Juve, Lana, and Manali, my companions throughout the Master's journey, for their endless encouragement, collaboration, and friendship. Sharing challenges, laughter, and progress together made this experience truly meaningful.

To all who contributed ideas, feedback, or moral support throughout this research, I really appreciate for the support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Generative artificial intelligence (AI) has rapidly advanced in recent years, with large language models (LLMs) and vision-language models (VLMs) achieving state-of-the-art performance across various tasks. These systems are now used not only to generate natural language or describing images and videos, but also to perform evaluative functions, such as rating, scoring, or judging outputs. As AI systems take on these evaluative roles, ensuring that they provide reliable and interpretable judgments becomes increasingly important for both research and practical applications. Yet, despite the impressive multimodal reasoning abilities of LLMs and VLMs, evaluating their performance remains a fundamental challenge. Traditional machine learning metrics such as accuracy, F1-score, BLEU [37], METEOR [15], or CIDEr [47] treat all items as equally difficult and assume that aggregate scores provide sufficient insight into model performance. While these metrics are useful for standardized benchmarks, they often fail to capture nuanced differences between models, overlook item-

level variation, and provide limited interpretability for how or why models succeed or fail on particular tasks. This limitation is especially evident in domains that require subjective judgments or multiple quality dimensions, such as captioning, comprehension, or accessibility.

This research is motivated by an applied accessibility challenge. In the context of YouDescribe [53], a crowdsourced platform where volunteers provide audio descriptions (AD) of online video content for blind and low-vision (BLV) users, the quality of the AD depends not only on accuracy, but also on a range of qualitative factors that determine whether a description truly meets the needs of BLV users. Since there are various aspects to be considered, the evaluation of quality has been identified as a recurring bottleneck. Human raters can provide this feedback, but the process is costly, subjective, and difficult to scale. These challenges highlight the need for scalable, interpretable, and rigorous methods of evaluating description quality, raising the question of whether AI systems themselves can serve as supplemental raters.

The central problem addressed in this thesis is how to evaluate the VLMs as raters. To refine this broader problem, the following guiding questions are posed:

- How can the ability of VLMs as evaluators, rather than generators, be measured?
- How can variation in item difficulty across tasks be incorporated into evaluation?
- How can alignment or divergence between VLM judgments and human ratings be identified?
- How can evaluation methods provide interpretability and validity in assessing model performance?

Addressing these questions requires a measurement-based approach that goes beyond conventional metrics. While AI research has primarily focused on generation, far less attention has been given to the evaluation capabilities of models. In contrast, psychometrics, particularly Item Response Theory (IRT) [18, 42], offers a robust methodology for measuring respondent ability and item difficulty on a shared scale, providing interpretable insights that traditional metrics cannot. Although IRT is widely used in psychometrics, it has not been systematically applied in machine learning evaluation. This thesis applies IRT as a means of bridging these perspectives, treating VLMs as "respondents" in evaluation tasks and assessing their performance alongside human raters. Moreover, a scalable Item-Person Map [27], also known as a Wright Map [27], is applied to provide interpretable visualizations that situate respondents and items together, making it possible to observe which models approximate human raters, where they diverge, and which items are inherently easier or harder to evaluate.

The adoption of IRT provides the methodological foundation for the following analyses. It allows model evaluation to move beyond accuracy-based metrics toward an interpretable measurement framework capable of comparing human and AI raters on a common latent scale. To explore this framework, three case studies were designed that progress in complexity and realism, each grounded in evaluation tasks related to accessibility:

1. **Image Caption Rating** (Baseline Task): VLMs were asked to rate image-caption pairs, with their evaluations compared to human ratings using the Validated Image Caption

Rating (VICR) dataset [33, 43]. This case serves as a foundational test of whether models can approximate human judgments at the five different level.

2. **Visual Reading Comprehension** (Comics as Proxy for Video): Comics (cite) provide a unique test case, combining sequential panels with embedded text. This task requires both local and global inference, paralleling the challenges of multi-frame video comprehension. By incorporating comics, the study examined whether VLMs can integrate across frames and modalities in a manner similar to human reasoning.

3. **Video Audio Description Rating** (Applied Accessibility Task): Full videos with ADs were evaluated across seven quality dimensions, including accuracy, timing, and appropriateness. This case represents the most complex and realistic scenario, directly tied to accessibility needs for BLV users. Here, IRT was extended to multi-dimensional modeling, reflecting the multifaceted nature of description quality.

From single frame captioning to sequential comprehension to full video description, each stage adds complexity while remaining aligned with the central motivation of evaluating AI as raters for accessibility-related tasks. Through a exploration of the case studies, the research demonstrates how psychometric methods can enhance both the validity and interpretability of AI evaluation. By grounding each study in accessibility-related tasks, the work not only addresses a critical application domain but also establishes a broader foundation for integrating measurement science into the assessment of AI systems.

Chapter 2

Related Work

This chapter situates the present research within four intersecting domains: (1) evaluation in machine learning research, (2) the recent emergence of AI-as-a-judge or AI-as-a-rater paradigms, (3) psychometric measurement methods and preliminary of IRT (4) evaluation practices in accessibility. Although each of these areas has developed largely independently, they converge in this thesis through the goal of establishing a rigorous, interpretable, and scalable framework for assessing VLMs as raters.

2.1 Evaluation in Machine Learning

Traditional Metrics and Benchmarks

Performance evaluation in machine learning traditionally relies on scalar metrics that summarize model accuracy over a dataset. Common measures include accuracy, precision,

recall, F1-score, and, for text generation tasks, automatic metrics such as BLEU [37], METEOR [15], ROUGE [23], CIDEr [47], and SPICE [4]. These metrics have been instrumental in benchmarking progress on datasets such as ImageNet [14], MS-COCO Captions [24], Flickr8k/30k [20], VQA [6], and MSR-VTT [51], enabling quantitative comparison across competing models. While useful for standardized leaderboards, these metrics share a simplifying assumption: every sample in the dataset is treated as equally difficult. The resulting aggregate scores conceal item-level variability and offer limited insight into why a model succeeds or fails. Moreover, many metrics rely on surface-level n-gram overlap or discrete accuracy, which poorly approximate human judgments of quality or coherence, especially for generative outputs such as captions or descriptions.

Human Evaluation: The VICR Dataset

Recognizing the limitations of purely automatic metrics, researchers introduced human-rated datasets such as Validated Image Caption Rating (VICR). VICR employs a novel five-level scale designed to capture multiple dimensions of caption quality (accuracy, completeness, local and global context), and it collected over 68,000 ratings from 113 participants over 15,646 image-caption pairs. Compared to earlier datasets like Flickr8k-Expert, VICR achieves significantly higher inter-rater agreement, suggesting greater reliability and consistency. However, despite its improvements, VICR remains constrained by the costs, time, and scale associated with human evaluation; it cannot be easily repeated for every new model,

dataset, or domain. In other words, while VICR moves evaluation beyond automated metrics, it still belongs to the human-in-the-loop paradigm, which faces scalability limitations. This suggests the need for AI-in-the-loop or measurement-based solutions that can scale reliably without losing interpretability or alignment with human judgment.

Limitations of Conventional Evaluation

Several limitations follow from existing evaluation paradigms. Traditional metrics ignore item difficulty and provide limited interpretability, while human-in-the-loop approaches, though richer, remain slow and resource-intensive. AI-in-the-loop evaluation promises scalability but lacks theoretical grounding and calibration. Together, these trends underscore the need for frameworks that are both interpretable and statistically principled, a gap this thesis addresses through the integration of IRT.

2.2 VLMs as Raters or Judges

LLMs and VLMs as Evaluators

A growing trend in recent research treats large models themselves as evaluators rather than solely generators. Frameworks, such as MT-Bench [7], Chatbot Arena [11], and ShareGPT4V [10], employ foundation models like GPT-4 or Gemini 1.5 to rate or rank outputs produced by other systems. This "AI-as-a-judge" paradigm offers scalability and consistency relative to

costly human annotation and has been adopted in reinforcement learning from AI feedback (RLAIF).

Despite these advantages, major challenges remain. Model-based evaluators can exhibit prompt sensitivity, positional bias, or preference drift depending on formatting. Without calibration, their ratings may not align with human criteria or ground truth, potentially reinforcing systematic biases rather than mitigating them.

Alignment and Human-AI Agreement

Recent studies attempt to quantify agreement between AI and human evaluators using rank-correlation statistics such as Kendall's τ [21] or Spearman's ρ [44]. While such correlations measure consistency at a global level, they provide little interpretability: a high τ does not reveal which items cause disagreement or how rating difficulty varies across models. Consequently, the field lacks a principled framework for analyzing evaluator ability and item complexity simultaneously which is an issue central to this thesis.

2.3 Preliminary of Item Response Theory

Foundations and Context

IRT is a statistical framework developed in the field of psychometrics to model the relationship between an individual's latent ability and their probability of correctly responding to

test items. In the measurement sciences, it has long served as the methodological backbone of large-scale standardized testing programs such as the Scholastic Aptitude Test (SAT), Graduate Record Examination (GRE), Program for International Student Assessment (PISA), and numerous professional-licensure exams. Its central premise is that a person's observed performance reflects not only their intrinsic ability but also the inherent difficulty and discriminative strength of each item.

Formally, IRT describes the probability that respondent j correctly answers or endorses item i as a logistic function of the difference between the respondent's ability (θ_j) and the item's parameters. This probabilistic modeling contrasts with classical test theory, which assumes equal weighting of items and treats measurement error as uniform across a scale. In IRT, measurement precision varies by ability level: the model captures where an instrument is most or least informative.

The Rasch Family and Logistic Models

The simplest and most interpretable member of the IRT family is the One Parameter Logistic (1PL), also kown as the Rasch Model [26]. Its form is

$$P(X_{ij} = 1|\theta_j, \delta_i) = \frac{e^{\theta_j - \delta_i}}{1 + e^{\theta_j - \delta_i}} = \frac{1}{1 + e^{-(\theta_j - \delta_i)}} \tag{2.1}$$

where

- $P(X_{ij} = 1)$ is the probability that respondent j provides a probability of success

response to item i,

- θ_j represents ability of respondent j, and
- δ_i represents the difficulty of item i

When plotted along a single latent dimension, items with higher δ are more difficult, while respondents with higher θ are more likely to succeed. Because both are placed on the same continuum, the Rasch Model yields a construct map of the measurement space that is invariant across test forms and populations.

Building upon the 1PL foundation, additional logistic models introduce further parameters which are not part of the Rasch Family Model:

- Two Parameter Logistic (2PL) model adds a *discrimination* parameter α_i, allowing items to differ in how sharpy they distinguish between high and low ability respondents.

$$P(X_{ij} = 1|\theta_j, \delta_i, \alpha_i) = \frac{1}{1 + e^{-\alpha_i(\theta_j - \delta_i)}} \tag{2.2}$$

- Three Parameter Logistic (3PL) Model adds a *guessing* parameter c_i, representing the lower asymptote of the curve which is the chance of success by guessing in multiple choice contexts.

$$P(X_{ij} = 1|\theta_j, \delta_i, \alpha_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + e^{-\alpha_i(\theta_j - \delta_i)}} \tag{2.3}$$

Dichotomous and Polytomous Extensions

Early IRT models assumed dichotomous scoring each response coded as correct (1) or incorrect (0). However, many real world assessments involve ordinal or graded responses, such as essay rubrics or Likert-type ratings. To accommodate these, IRT was extended to polytomous models, which estimate multiple ordered thresholds per item rather than a single difficulty value. Prominent examples include:

- Partial Credit Model (PCM) [26, 27]
- Rating Scale Model (RSM) [5]
- Graded Response Model (GRM) [41]

Polytomous IRT enables the analysis of nuanced rating data, estimating how respondents transition between adjacent score categories. These models are particularly relevant for research domains that rely on subjective or multi-level judgments, such as human ratings of captions, essays, or audio descriptions, because they capture the probabilistic structure of ordinal decisions rather than collapsing them into binary outcomes.

A major advantage of the Rasch 1PL model and its extensions lies in interpretability. In psychometrics, item and respondent parameters are visualized through an Item-Person Map [27], also known as a Wright Map, which displays both on a shared logit scale to illustrate where items challenge respondents and where they perform reliably. This interpretive clarity linking numerical estimates to a meaningful visual representation makes IRT particularly appealing for domains that require transparent evaluation.

Although IRT has become a standard in educational measurement, it has rarely been applied in machine learning evaluation. In this thesis, these psychometric principles are adapted to assess VLMs acting as raters. By treating models as "respondents" and evaluation items, IRT can provide a principled framework for estimating model ability and item difficulty within a common probabilistic space. This conceptual foundation underlies the methodological approach detailed in the next chapter.

2.4 Accessibility and Evaluation

Audio description (AD) enhances access to visual media for blind and low-vision (BLV) audiences by narrating key visual details such as actions, settings, and character interactions. Professional organizations have established comprehensive standards for ensuring quality: the Described and Captioned Media Program (DCMP) [16] provides educational guidelines, and the National Center for Accessible Media (NCAM) at WGBH has defined conventions for broadcast and film [17]. Commercial providers such as 3Play Media [1] integrate AD into structured accessibility workflows involving multi-stage review processes that include script writing, peer review, editorial checks, and pilot testing with BLV users [36]. While these processes make professional AD the gold standard for quality and consistency, they are labor-intensive, descriptive rather than quantitative, and dependent on expert oversight.

To extend access beyond professional production contexts, volunteer-driven platforms such as LiveDescribe [9], Rescribe [38], and YouDescribe [53] adapt professional guidelines

for non-expert contributors. These crowdsourced approaches have significantly expanded AD availability, yet their outputs often vary in accuracy, timing, and clarity. Studies show that volunteer describers, particularly novices, frequently produce descriptions that are misaligned with audio, omit key visual details, or lack appropriate pacing [32, 38]. On average, volunteer-generated AD achieves only about 60% of professional-quality benchmarks and is subject to inconsistencies in loudness, completeness, and delivery strategies. The rapid growth of online video platforms such as YouTube, TikTok, and Instagram further amplifies the challenge: the sheer scale of content far exceeds the capacity for professional or manual review, underscoring the need for scalable methods of quality evaluation.

In parallel, recent advances in artificial intelligence have introduced AI-generated AD, where multimodal models automatically generate descriptions based on video input. Early systems relied on rule-based multimodal narration pipelines [49], while recent approaches employ LLMs and VLMs for scene-aware captioning [12, 46, 52]. These methods enable scalable generation but frequently produce inaccurate or cluttered descriptions that misidentify objects, hallucinate entities, or fail to synchronize with the corresponding audio [8, 40, 49]. Despite achieving high scores on captioning benchmarks, AI-generated AD often diverges from human perceptions of accessibility-critical qualities such as relevance, coherence, and pacing [40]. Consequently, both volunteer and AI-generated descriptions suffer from the absence of rigorous, interpretable evaluation standards that account for the multidimensional nature of AD quality.

Efforts to evaluate AD quality have largely followed two paths: automatic metrics and

human-centered assessments. Automatic metrics, such as BLEU [37], METEOR [15], CIDEr [47], and SPICE [4], enable objective and reproducible comparisons but privilege surface-level similarity and overlook temporal and contextual aspects critical to accessibility. Human-centered evaluations, typically involving BLV or sighted participants, provide richer qualitative feedback but remain limited by recruitment challenges, fatigue, and inconsistent annotation quality [28, 31, 50]. More recent hybrid frameworks, such as VideoA11y [22], combine accessibility-informed dimensions (descriptive, objective, accurate, clear) with conventional NLP metrics to yield a more comprehensive view of AD quality. However, these studies are resource-intensive and often restricted to short clips from benchmark datasets such as VALOR [25], YouCook2 [54], or VATEX [48]. As the field continues to balance scale, validity, and interpretability, establishing rigorous, scalable methods for evaluating both volunteer and AI-generated AD remains an urgent challenge within accessibility research. The challenges outlined here underscore the need for scalable and transparent evaluation methods that extend beyond surface-level metrics. The next chapter details a psychometric framework designed to provide interpretable, statistically grounded insights into model and human evaluation behavior.

Chapter 3

Analytical Framework

This chapter describes the analytical framework and modeling procedure applied consistently across the three case studies presented in this thesis. Each study adopts a common psychometric foundation using IRT to model the relationship between rater ability and item difficulty based on ordinal evaluation data.

Rasch Model

Across all case studies, the same core methodology is used. Each VLM and each human rater is treated as a respondent, and each evaluation instance is treated as an item.

The analytical goal is to estimate two latent parameters jointly:

- **Rater ability (θ):** the inferred evaluative competence of a rater
- **Item difficulty (δ):** the relative challenge of evaluating a given item correctly or consistently.

A 1PL or Rasch model is used as the principal formulation in all analyses. The 1PL model provides a stable and interpretable measurement scale, assuming equal discrimination across items while emphasizing comparability among raters. More complex models such as the 2PL or 3PL models introduce additional parameters to estimate item discrimination and guessing behavior, respectively. While these extensions can capture item-specific variability, they complicate interpretation and comparability when discrimination values differ across items. Moreover, unequal discrimination precludes direct visualization through Wright Maps. For this reason, the 1PL (Rasch) model is preferred in this thesis, as it maintains interpretability, stability, and visual coherence across all case studies.

Moreover, many rating tasks in this research use ordinal response categories rather than binary outcomes. For example, both human and AI raters evaluated items on a 1-5 scale, where higher values indicate greater quality or alignment. To apply IRT in this context, these raw ordinal ratings must first be linked or compared to a benchmark or ground truth judgment, established through expert or consensus ratings. Without this reference, differences between model and human judgments cannot be interpreted probabilistically. In practice, however, converting ordinal ratings directly into binary outcomes (e.g., "correct" vs. "incorrect") would discard valuable information about near agreement. For instance, when a model assigns a score of 4 while the expert benchmark is 5, the model's judgment is not fully incorrect; it remains partially aligned. Treating such cases as errors under a dichotomous Rasch model would erase meaningful evaluative signal and overstate disagreement.

Partial Credit Model

PCM [27] generalizes the Rasch model for ordered, multi-category data by assigning partial credit to responses that are close to the benchmark rating. Rather than modeling only binary success, PCM evaluates how far a response deviates from the ground truth, capturing partial and complete agreement. In this thesis, a partial credit scoring scheme was constructed based on the distance between each rater's score (human or AI) and the expert benchmark for a given item:

- **2 (Exact Agreement):** The respondent's rating matched the ground truth rating exactly.
- **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 scoring transformation produces a 0–2 ordinal scale, where higher values represent stronger alignment with expert judgment. By adopting this scheme, the model recognizes that near misses (adjacent agreement) still convey partial evaluative accuracy, instead of collapsing them into full errors such that it can provide a more nuanced and continuous picture of rater performance.

Mathematically, the PCM models the probability that respondent j achieves score k on item i as:

$$P(X_{ij} = k \mid \theta_j, \delta_{i1}, \ldots, \delta_{im}) = \frac{\exp\Big(\sum_{s=0}^{k}(\theta_j - \delta_{is})\Big)}{\sum_{t=0}^{m}\exp\big(\sum_{s=0}^{t}(\theta_j - \delta_{is})\big)}. \tag{3.1}$$

Here, θ_j represents the respondent's latent ability (i.e., evaluative proficiency), while δ_{is} denotes threshold parameters separating adjacent categories. Each threshold represents the point at which a rater is equally likely to assign either of two neighboring levels of agreement. For example, particularly with 0-2 scale of PCM, each item is associated with two thresholds. Threshold 1 represents the point where a respondent (or a rater) 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 (or a rater) 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.

Using PCM in this way allows the model to estimate both how well each rater aligns with expert benchmarks and how difficult each item is to rate consistently. Furthermore, person and item fit statistics along with item correlation relationships can be generated that help determine the confident level of person proficiency estimates and well-fit of items. The Mean Square (MNSQ) fit statistic is a quantitative way to evaluate the validity of the Wright Map. The acceptable range is 0.75 to 1.33, with 1.0 representing the ideal value. Items or Respondents with a fit statistic above 1.33 indicate under fit which shows excessive randomness.

Interpreting the Item–Person (Wright) Map

This Sub-Chapter provides a common reading guide for the Item-Person Maps, or the Wright Maps used throughout all 3 case studies. A Wright Map places respondents (humans and VLMs) and items on a shared latent scale, enabling direct visual comparison of who tends to agree with expert judgments and which items are easier or harder to evaluate. Establishing this shared vocabulary upfront allows later chapters to focus on substantive findings.

Although visual styles vary slightly across figures of each case study, the map always show the common elements:

- **Respondent Distribution:** A stacked or histogram-like distribution of respondents (humans and VLMs) displays how abilities are spread along the logit axis. Dense regions indicate many raters with similar ability; tails indicate especially strong or weak raters relative to the benchmark. This "at a glance" view complements individual θ estimates reported elsewhere.

- **Items with Thresholds:** Each item appears with ordered thresholds on the same scale. With three ordered categories (0, 1, 2), there are two thresholds:

  - **Threshold 1:** is the point where a respondent is equally likely to move from distal agreement (0) to at least adjacent agreement (1 or 2)

  - **Threshold 2:** is the point where a respondent is equally likely to move from (0 or 1) to exact agreement (2).

  These thresholds situate the "difficulty" of achieving higher agreement levels and enable fine-grained comparisons across items.

- **Cutpoints:** A doted horizontal line that signifies theoretical or substantive regions of interest. They are a visual indicator showing certain items correspond to certain regions. By glancing, it is easier to see which respondents are associated with which region and where the test provides most information.

Vision Language Models

To examine how accurate AI-generated ratings are and how accuracy compares with that of human generated ratings, each case studies was conducted using a diverse set of VLMs. For Case Studies 1 and 2, the following models were used:

- **SmolVLM-Instruct** [39] A compact, 2.25B-parameter model by HuggingFace optimized for instruction following and fast inference. Released November 18, 2024.
- **Phi-3.5-Vision-Instruct** [30] A 4.2B-parameter vision–language model by Microsoft trained on synthetic and filtered public data. Released August 20, 2024.
- **Qwen2-VL-7B-Instruct** [45] A 7B-parameter model by Alibaba Cloud optimized for multimodal instruction following. Released October 3, 2024.
- **Gemini-1.5-flash** [2] A proprietary sparse Mixture-of-Experts (MoE) model by Google; parameter count undisclosed. Released May 14, 2024.
- **Llama3.2-11B-Vision-Instruct** [29] An 11B-parameter model by Meta designed for multimodal instruction tasks. Released September 25, 2024.

- **GPT-4o family** [35] Includes GPT-4o (omni), GPT-4o-mini, and GPT-4-turbo variants by OpenAI. "Omni" models process text and image embeddings jointly; "mini" and "turbo" emphasize efficiency and speed. GPT-4o-mini was released July 18, 2024.

For Case Study 3, limited amount of VLMs were able to use videos as inputs, hence only three models listed below were used:

- **Qwen2.5-VL** [3] An open-source model by Alibaba Cloud supporting image and video understanding through frame-level encoding and temporal attention. Released September 19, 2024, it features an enhanced visual encoder for short video reasoning.
- **Gemini 1.5 Pro** [19] A closed multimodal transformer by Google DeepMind capable of handling video, image, audio, and text within a shared context. Built on a sparse Mixture-of-Experts (MoE) architecture for long-context and temporal reasoning. Released February 15, 2024.
- **GPT-4o** [34] A unified multimodal model by OpenAI that directly interprets video, audio, image, and text inputs. It aligns sampled frames with temporal context embeddings for fine-grained scene comprehension. Released May 13, 2024.

To capture a broad variance in model performance, the selection covered a spectrum ranging from lightweight models to state-of-the-art large-scale systems. Each model is prompted under standardized evaluation instructions on each case study to produce rating scores comparable to human responses. Furthermore, default hyperparameters such as temperature, batch size, These ratings are used to construct PCM and then entered into the response matrix used for IRT estimation.

Chapter 4

Case Study 1: Image Caption Rating

Image captions play a crucial role in bridging visual content and linguistic expression, enabling systems to communicate what they see in ways that support accessibility and understanding. This first case study establishes a baseline for evaluating whether VLMs can approximate human judgment in rating image-caption pairs that have traditionally relied on human evaluation. The focus is on comparing how models perceive caption quality and how well a description captures the main content of an image, conveys context, captures salient features, and reflects overall coherence. Because this task lies at the intersection of visual understanding and linguistic reasoning, it offers a clear and interpretable setting to observe model judgment in a structured way.

4.1 Study Design

Dataset and Rating Scale

To explore how well VLMs can approximate human judgment on image-caption quality, two datasets were investigated: Flickr8k [20] and VICR [33, 43]. These two datasets represent different stages of development in caption evaluation research. Flickr8k is one of the earliest benchmarks, and VICR is more recent, psychometrically validated framework for human caption assessment.

1. Is unrelated to the image.

2. Is somewhat related to the image.

3. Describes the image with minor errors.

4. Describes the image without any errors.

Figure 4.1: Flickr8k rating scale from 1 to 4

The Flickr8k dataset is a classic benchmark for caption generation and evaluation. It contains 5,822 image-caption pairs, each rated by human annotators using a four-point ordinal scale (Figure 4.1). While useful for early comparative analysis, such as measuring rank correlations between human and model judgments through Kendall's τ, the Flickr8k dataset offers limited granularity in both rating distribution and scale design. The four-level rubric captures only surface-level correctness rather than nuanced semantic coherence or contextual interpretation. Consequently, Flickr8k was not employed in the IRT analysis, but served as

1. Objects are incorrectly identified. The caption gives the wrong idea about what is happening in the image.
2. Objects are partially correctly identified with some errors, but the caption is accurate enough to give an idea of what is happening in the image. The caption identifies most of the objects but might not identify everything. There is no interpretation of what anything means.
3. Relevant objects are correctly identified. The caption describes what is seen but not where objects are in space. There is no description of the overall setting and no interpretation of an event.
4. Objects and/or a general scene and/or an action are correctly identified but not every element is completely identified. The caption describes what is seen and where things are in space. There is no interpretation of an event.
5. Objects, a general scene, and actions are correctly identified if present in the image. The caption describes what is seen and where things are in space. Interpretation of overall setting and/or event is included.

Figure 4.2: VICR rating scale from 1 to 5

a preliminary reference point for understanding how VLMs perform on traditional caption rating benchmarks.

The VICR dataset represents a major advancement in caption quality assessment. Developed to model human perception of caption quality more systematically, VICR consists of 15,646 image-caption pairs, each rated by 3 to 7 human participants on a five-level scale (Figure 4.2). The VICR rubric captures both local properties, such as object identification and spatial accuracy, and global properties, such as event interpretation and contextual coherence. Its balanced rating distribution and psychometric calibration make it well suited for latent-trait modeling through IRT.

For this case study, rather than using the full VICR dataset, a subset of 25 items was

used (see Appendix .1). This is because the IRT framework requires an expert benchmark as a reference, against which model ratings are compared, so that alignment with ground truth can be measured on a common latent scale; explained detail in "Construct Map" Sub-Chapter. These items were independently constructed by the same expert group that developed the VICR dataset. Each item represents a single image paired with five captions corresponding to levels 1 through 5 on the VICR scale, effectively forming a controlled mini-benchmark for rating behavior analysis. These curated items were originally introduced in the "Rating Game" [33, 43], a human evaluation experiment designed to validate the VICR scale through targeted examples of varying caption coherence. Total 132 human ratings were able to be collected and used in IRT analysis.

Prompt Design

To collect the VLMs' evaluations or ratings, structured prompt was built that guided each model through the image-caption evaluation task. Considering that prompts are highly subjective and can significantly influence model performance, the prompts were designed following the "Hierarchical Prompt" methodology introduced in the ShareGPT4V [10] template as a foundational framework. Figure 4.3 illustrates the hierarchical prompt structure used to guide model evaluation.

To ensure the VLMs effectively perform, a specific role was assigned, treating them as a "character" within the prompt. The prompts are structured to optimize the model's

You are an expert at determining how good a description is for an image. You are able to rate the given description and image on a scale of 1 to 5, corresponding to how well the description fits the image. Using the following rubric of rating levels:

<<VICR Scale (from Figure 4.2)>>

Please rate the following description on how well they describe the given image. Just the number (ratings from 1 to 5).

Figure 4.3: Structured prompt for VLMs to generate ratings on image-caption pairs.

performance in assessing caption quality. First, the role of the model was defined, explicitly guiding it toward the task of rating captions. Then, the rating scale as a reference framework to standardize its evaluations was provided. Lastly, to solely retrieve ratings, each model was prompted to generate only the numbers, and provide the image and associated caption. Each VLM was provided with the same prompt and importantly, each rating was generated in isolation – the models did not retain any context or memory of previous inputs, ensuring that each rating was produced independently without carryover effects.

Partial-Credit Scoring Construction

Applying the IRT framework to image-caption rating requires an additional preparation step to make model outputs commensurable with human judgments. Each model's raw rating is first aligned to an expert benchmark and then transformed using the PCM. Concretely, model ratings are recoded on a 0-2 ordinal scale based on their distance from the expert level(rating). As mentioned in Chapter 3 under "Partial Credit Model" Sub-Chapter,

2 represents exact agreement, 1 for adjacent agreement (± 1 point), and 0 for distal agreement (≥ 2 points). This is to preserve near-miss information that would otherwise be lost under a dichotomous scheme, yielding a response matrix that is informative for latent-trait estimation.

4.2 Result and Discussion

VLMs	θ	**Percentile**
gpt-4-turbo-2024-04-09	-0.21015	33%
gpt-4o-2024-05-13	**0.52807**	**89%**
gpt-4o-2024-08-06	0.01775	54%
gpt-4o-2024-11-20	**0.52807**	**89%**
gpt-4o-mini	0.01775	54%
Gemini_1.5_flash	-0.09786	41%
Llama3.2-11B-vision_instruct	-0.53158	16%
SmolVLM-Instruct	-0.73605	12%
Phi-3.5-vision-instruct	-0.31962	24%
Qwen2-VL-7B-Instruct	0.01775	54%

Table 4.1: VLMs' Proficiency score (θ) and Percentile.

Across the 10 VLM respondents, person abilities span roughly -0.74 to $+0.53$ logits, with several systems clustered near 0. Specifically, GPT-4o-2024-11-20 and GPT-4o-2024-05-13 had the highest θ out of all VLMs; these top two VLMs also performed better than or equal to 89% of 142 respondents (132 humans and 10 VLMs). There are 4 other VLMs, Qwen2-VL-7B-Insturct, Gemini-1.5-Flash, GPT-4o-mini, and GPT-4o-2024-08-06 that scored close to 0, suggesting that those VLMs are as good as the average human respondent. Moreover, GPT-

4-turbo-2024-04-09, Phi-3.5-Vision-Instruct, Llama3.2-11B-Vision-Instruct, and SmolVLM-Instruct performed below the average respondent suggesting they are not performing well at the image-caption rating task, relative to the other respondents.

Wright Map Interpretation

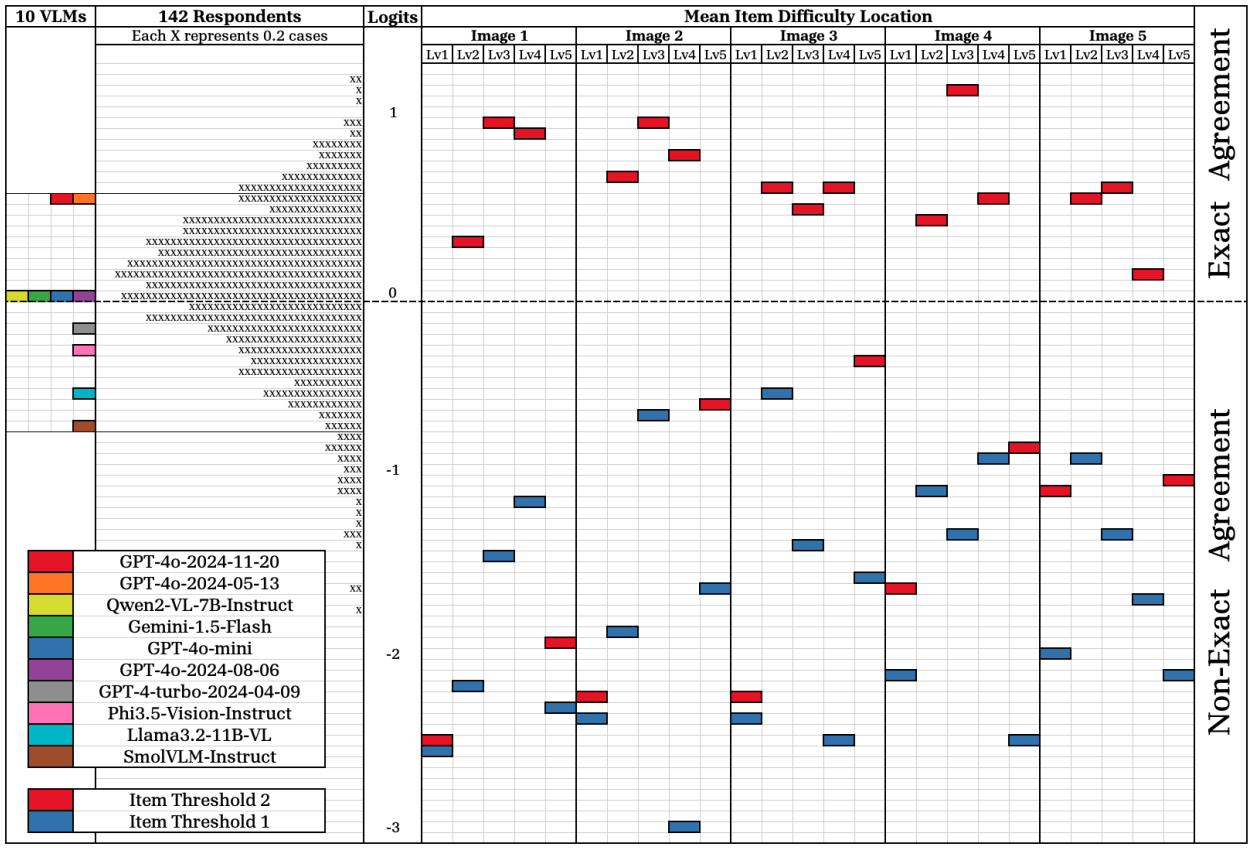


Figure 4.4: Wright Map of VICR study showing the 142 respondent proficiency (θ), and 25 item difficulty on the same logit. "Non-Exact Agreement" and "Exact Agreement" are the region of interest from the cutpoint.

The Wright Map places all 142 respondents, 132 humans and 10 VLMs, on the same logit scale as the items' partial-credit thresholds. As shown in figure 4.4, from left to right, it

represents where 10 VLMs are located, all respondents distribution, 25 items (Appendix .1 associated with 2 thresholds each and the cutpoint on the same logit scale. Additionally, to give better idea of how to interpret the respondent performance, 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.

From the person-level observation, the VLMs concentrate near the center of the scale, with a few models above and below the average human cluster. This indicates 10 VLMs perform average to slightly above average alignment with the benchmark. This suggests that their performance is not perfect, but also not uniformly worse than human performance. Moreover, they are spread out and located both in Exact Agreement and Non-Exact Agreement regions from the cutpoint which points out that this IRT analysis is informative.

Regarding the item-level observation, the interesting pattern was found. Noticed that the item thresholds, both 1 and 2, associated with extreme categories of rating 1 and 5 lie below the respondent mean where as those item thresholds associated with mid categories of rating 2, 3, and 4 place above. This implies that achieving exact agreement on the items with rating 1 and 5 are much easier while it is harder for the items in the mid range ratings. In other words, respondents readily agree when captions are obviously poor or strong, but have hard time rating when the quality of captions are mixed or context-dependent.

4.3 Limitation and Future Work

VLMs can be used as a viable option for image–caption rating on this task, but a human remains necessary in the loop because the models are not uniformly reliable, especially in the mid-range cases where exact agreement is hardest to achieve. The Wright Map shows that models and humans overlap around the center of the scale; this is desirable for targeting, but it also means that borderline items remain the primary source of disagreement. A practical implication is to allocate work asymmetrically: allow VLMs to triage clear successes and clear failures, and reserve ambiguous cases for human adjudication. This hybrid arrangement preserves efficiency without sacrificing the validity of final judgments.

Methodologically, several elements limit generality and suggest extensions. The item pool was deliberately small (25 items) to provide a proof-of-concept map; scaling the pool while balancing category representation across 1-5 would increase separation reliability and reduce dependence on a handful of difficult thresholds. A larger, stratified pool should include diverse image types so that difficulty is not dominated by a narrow content band. Model diversity is another constraint. Although the respondents include multiple families, the set is not large enough to support strong claims about architecture-level differences. Taken together, these aspects would convert this proof of concept into a robust measurement instrument.

Chapter 5

Case Study 2: Visual Reading Comprehension

This study focuses on visual reading comprehension, extending the evaluation framework introduced in the previous chapters to a multimodal reasoning task. While the image caption rating study (Chapter 4) assessed how VLMs interpret visual–textual alignment at the descriptive level, this case examines their ability to understand and infer meaning across sequential narratives that combine images and text.

5.1 Study Design

This study employed a reading comprehension task adapted from Blum et al. (2024) to examine integrative inferential reasoning across different narrative modalities. Three moral

narratives were used as stimuli: *Friendship*, *Stealing*, and *Lying*. Each narrative has two qeuivalent modalities: comic + text (multimodal; text is embedded to the comic panels) and text-only (verbal). The comics followed Cohn's visual narrative grammar framework [13], incorporating sequential panels that represented key story elements such as establishing, initiating, peak, and resolution scenes. The accompanying text was identical in both conditions, written at a third-grade readability level, based on Lexile Readability Scores, to minimize text complexity as a confounding variable.

Dataset

A total of 130 middle and high school students (ages 11–17) from schools in Northern California participated in the original human study. Participants were randomly assigned to one of two conditions (modalities), comic + text or text-only and responded to four open-ended inferential questions following each story. These questions, developed from established reading comprehension taxonomies, included one motivational inference question specific to each narrative and three general probes assessing higher-order reasoning:

1. **Motivational Inference (story-specific "Why" question):**
    - Why did the red-headed boy keep the cat?
    - Why did Billy keep the lady's wallet?
    - Why doesn't Sarah tell the teacher that Rachel is looking at another student's answers?
2. **Meta Reasoning (follow-up) Question:** What made you think of that answer?
3. **Evaluative Inference Question:** What lesson could someone learn from this story?
4. **Meta Reasoning (follow-up) Question:** What made you think of that answer?

Each participant therefore produced twelve responses (4 questions $\times$ 3 narratives), which were qualitatively coded to reflect increasing levels of inferential integration: "I don't know", "Local", "Global", and "Local + Global". These categories were subsequently assigned ordinal values of 0-3 to support IRT analysis under the PCM.

Response Evaluation and Coding Scheme

To convert open-ended responses into analyzable ordinal data, each answer was evaluated using a rubric grounded in Integrative-Inferential Reasoning (IIR) and applied through a structured classification protocol. The evaluation process relied on a rule-based system prompt designed to emulate expert raters who determine whether a response reflects local, *global*, or mixed inferential reasoning. For every question, the evaluator ignored repeated phrasing from the prompt, segmented the remaining text at conjunctions, and analyzed each sub-sentence individually. A sub-sentence was marked local if it referenced information explicitly stated or visually depicted in the story, and *global* if it required external or cultural knowledge extending beyond the text. When all sub-sentences were local, the overall classification was 1; when all were *global*, it was 2; and when both types appeared, it was 3. Responses containing no meaningful inference (e.g., "I don't know") were coded as 0. These four ordered categories (0–3) formed the ordinal response scale for the PCM analysis, enabling estimation of item difficulty and respondent ability on a shared logit scale.

Prompt Design

The same set of inferential questions used with human respondents was administered to 10 VLMs to examine their comprehension ability across narrative modalities. The model lineup was identical to the one used in the image caption rating study. While all models were evaluated for multimodal compatibility, not every system successfully processed both modalities. Nine models were tested under the comic + text condition (all except Llama3.2-11B-Vision-Instruct, which failed to produce coherent multi-image outputs), whereas eight models were tested under the text-only condition (excluding both Llama3.2-11B-Vision-Instruct and SmolVLM-Instruct, as the latter required at least one image input).

Each model received the same four inferential questions for all three narratives, following a structured prompt described in Appendix .1. Additionally, to reduce overly verbose responses that are typical of large models, we applied an explicit behavioral constraint in the system prompt:

> *"You are a middle school or high school student. Please answer the following questions based on the given panels in one sentence."*

This framing was intended to approximate the linguistic style and cognitive scope of the adolescent participants in the human dataset, thereby making the cross-group comparison more interpretable under a common scoring framework.

Once the responses were collected, they were evaluated using the same evaluation strategy described earlier for human data. Each answer was scored on a four-point ordinal scale: 0 (I don't know), 1 (Local), 2 (Global), and 3 (Local + Global), using the structured inference-

type evaluation rubric. The resulting coded responses were then incorporated into the same PCM framework as the human data, allowing direct comparison of human and VLM latent abilities on a shared logit scale.

5.2 Result and Discussion

VLM Respondent	**Comic + Text (θ)**	**Text-Only (θ)**
GPT-4-Turbo (2024-04-09)	**1.55903**	0.53525
GPT-4o-mini	1.00795	0.38988
GPT-4o (2024-11-20)	0.68559	**1.18214**
GPT-4o (2024-08-06)	1.18214	1.00795
GPT-4o (2024-05-13)	0.38988	0.53525
Phi-3.5-Vision-Instruct	1.00795	0.53525
Qwen2-VL-7B-Instruct	0.84269	0.68559
Gemini 1.5 Flash	1.00795	0.10637
SmolVLM-Instruct*	-1.25288	N/A

Table 5.1: 9 VLM respondents Proficiency or Ability (θ) with two modalities (Comic + Text, and Text-Only). *SmolVLM-Instruct was only able to response with image input.

Unlike the previous case studies, where higher scores reflected closer alignment with human benchmark ratings, this analysis measured each respondent's ability to produce more integrative inferences; that is, to progress from literal, text-bound answers toward global reasoning that combines contextual and cultural understanding. In the Partial Credit Model, higher person logits (θ) therefore indicate greater proficiency in generating Local + Global responses rather than mere accuracy.

As shown in Table 5.1, GPT-4-Turbo (2024-04-09) achieved the highest proficiency in the comic + text modality (θ=1.56), demonstrating the strongest multimodal reasoning

among the tested models. GPT-4o (2024-11-20) reached the highest score in the text-only modality (θ=1.18), suggesting advanced inferential reasoning even without visual support. The lowest proficiency was observed in SmolVLM-Instruct (θ=–1.25), which was expected given its limited parameter size and constrained visual processing capability. Interestingly, Gemini 1.5 Flash recorded the lowest θ in the text-only modality (θ=0.11) despite performing competitively in comic + text (θ=1.01), indicating that its comprehension strength relies heavily on multimodal cues rather than linguistic reasoning alone.

Additionally, VLM fit statistics all fell within the acceptable range of 0.75 to 1.33, confirming that none of the VLMs produced erratic or inconsistent patterns (see Appendix .2). Slightly lower fit values (e.g., GPT-4-Turbo comic = 0.55) suggest stable over fit, meaning these models consistently generated higher-level inferences rather than random fluctuations. Collectively, these results show that larger multimodal architectures tend to achieve more integrative reasoning, and that visual context particularly benefits high-capacity models while offering less advantage, or even confusion to smaller, lightweight systems.

Wright Map Interpretation

Figure 5.1 presents the Wright Map displaying the joint distribution of respondents (humans and VLMs) and item thresholds across the four comprehension questions for each of the three narratives (*Friendship*, *Stealing*, and *Lying*). Each "X" on the left indicates respondents' locations on the latent scale with text (text-Only) and visual (comic + text),

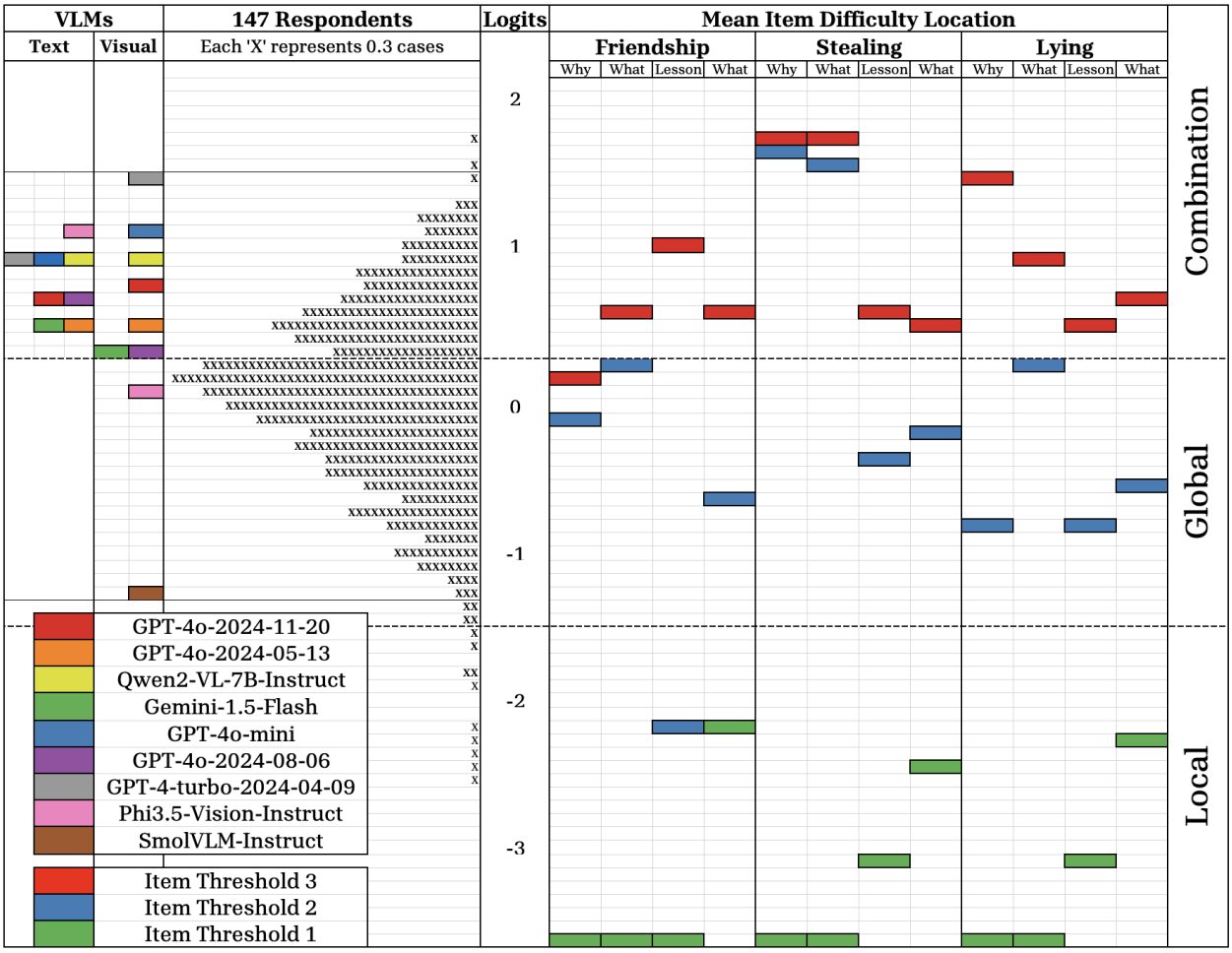


Figure 5.1: Wright Map of Visual Reading Comprehension study showing the 147 respondent proficiency (θ), and 12 item difficulty on the same logit. "Non-Exact Agreement" and "Exact Agreement" are the region of interest from the cutpoint.

while colored bars on the right represent the three category thresholds for each item under the PCM (0–3 scale). The thresholds correspond respectively to transitions from *Local* to *Global* (Threshold 1), from *Global* to *Combination* (Threshold 2), and the highest level of *Combination* (Threshold 3).

Overall, VLM respondents are located slightly higher on the logit scale than the mean of human respondents, indicating stronger tendencies toward *Global* and *Combination* (*Local*

+ Global) inference. This pattern is expected, as VLMs seldom produced "I don't know" or purely literal answers; their outputs typically contain inferential or interpretive reasoning even when minimal prompts were provided. The distribution shows two main rater clusters, one near the *Global* region and another around *Combination*, reflecting how both humans and VLMs naturally divided between moderate and advanced levels of inferential integration.

On the item side, the "Why" questions (the first inferential probes) were consistently more difficult, with higher thresholds required to reach the *Global* or *Combination* regions. This effect was most pronounced in the *Stealing* story, where many responses from both humans and VLMs remained tied to surface-level details ("He wanted to buy a red bike", "He took the wallet.") rather than extending to moral or psychological inferences such as "He is greedy" or "He is mean." By contrast, the *Friendship* story exhibited the lowest mean item difficulty, with several respondents, especially VLMs, achieving *Global* or *Combination* levels more readily. This likely reflects the simpler moral structure of that narrative and clearer emotional cues embedded in the panels.

Across all three stories, the Lesson ("What lesson could someone learn from this story?") and Reasoning ("What made you think of that answer?") questions were systematically easier, showing thresholds clustered in the upper *Global* and *Combination* regions. These items inherently encouraged moral generalization and reflection beyond the text, aligning closely with the conceptual definition of Global inference. The Friendship and *Lying* stories, in particular, prompted more consistent high-level responses, while Stealing remained challenging for all but the most capable models (e.g., GPT-4-Turbo and GPT-4o-2024-11-20).

Taken together, the Wright Map illustrates that higher latent ability corresponds to a stronger capacity for moral and integrative reasoning. Large multimodal models, which tended to occupy the upper end of the scale, demonstrated facility in producing both text-based and contextually enriched inferences. Smaller models, in contrast, remained constrained to local coherence and rarely reached the *Combination* region. These findings reinforce the interpretive role of IRT in distinguishing qualitative reasoning levels showing that progress along the logit continuum reflects a shift from literal recall to abstract moral understanding.

5.3 Limitation and Future Work

While this study provides an initial examination of visual reading comprehension through IRT, several limitations should be acknowledged. First, the dataset was relatively small, consisting of only three narratives (Friendship, Stealing, and Lying), each paired with four inferential questions. This yielded a total of twelve items, which constrained the range of item difficulty and limited the statistical precision of the model estimates. Expanding the dataset to include additional stories, genres, and question types would enhance construct coverage and allow for more robust inferences about model and human variability.

Second, the current VLM evaluation relied on presenting each comic as a sequence of individual panels. While this approach aligns with how humans naturally read visual narratives, it may also fragment contextual information that the models could otherwise integrate

if the entire comic were processed as a single image. Future work could therefore compare these two input strategies, panel-by-panel versus holistic full-comic input, to examine how spatial continuity and contextual scope influence inferential reasoning.

Further extensions could also incorporate models fine-tuned on narrative understanding or moral reasoning tasks, enabling deeper insight into how VLMs internalize abstract social cognition. Combining a larger, more diverse corpus with varied input modalities would yield a richer IRT framework capable of mapping the evolving comprehension capacities of multimodal systems.

Chapter 6

Case Study 3: Video Audio Description Rating

This third case study examines whether VLMs can evaluate the quality of AD for long-form or full video contents. While previous chapters focused on a single image and short visual narratives with static images, this study extends the analysis to multimodal, time-based media where narration, dialogue, and visual events must be synchronized. The motivation stems from the growing use of VLMs in accessibility contexts, where they are increasingly applied not only to generate but also to assess descriptions for BLV audiences.

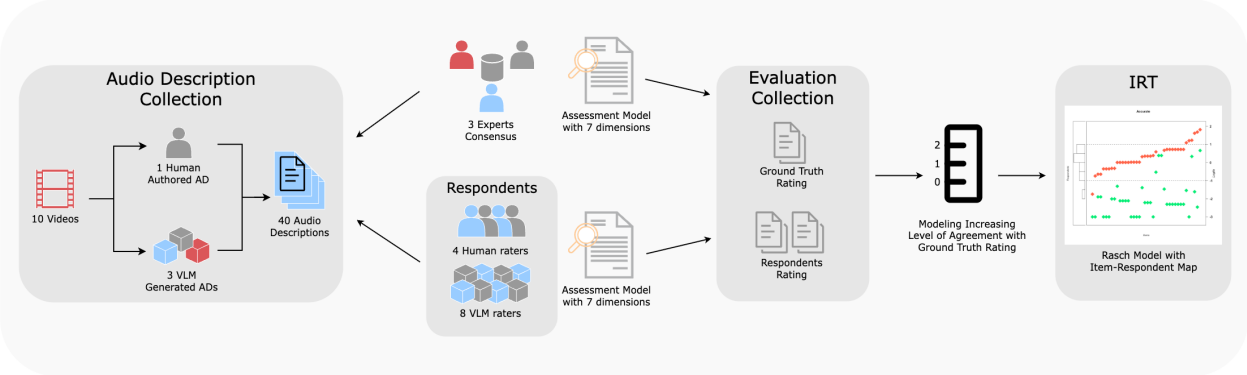


Figure 6.1: Conceptual diagram of the major components in the evaluation framework.

6.1 Study Design

Assessment Framework

The audio video description assessment framework was designed through close consultation with accessibility specialists experienced in creating and evaluating AD for BLV users. Their insights emphasized that the quality of AD depends on two equally critical dimensions: the content of what is described and the formatting of how and when descriptions are delivered.

Existing evaluation efforts in accessibility research have largely emphasized content accuracy while overlooking the delivery aspects that affect practical usability. In professional practice, however, inappropriate pacing or poorly timed narration can undermine even the most accurate descriptions. To address this gap, the framework formalizes seven dimensions of AD quality, five related to content and two related to formatting. They are adapted from the DCMP [16] and NCAM [17] professional guidelines.

Content Dimensions:

- **Accurate**: Descriptions are factually correct and error-free.
- **Prioritized**: Information critical for comprehension and enjoyment should be emphasized, while less important details are minimized.
- **Appropriate**: Language should be suited to the target audience, maintaining simplicity and conciseness.
- **Consistent**: Terminology, tone, and pacing should align with the program's style and remain uniform throughout the narration.
- **Equal**: Descriptions should preserve the program's meaning and intent for equitable access without distortion or bias.

Formatting Dimensions:

- **Strategic Use of Delivery 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.
- **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.

Each audio description was rated on a five-point ordinal scale (1 = lowest to 5 = highest) along these seven dimensions. Raters were provided with full documentation, including definitions, examples, and criteria corresponding to each scale point (see Appendix .7). The framework unifies long-standing professional standards with empirical evaluation methods, enabling systematic and reproducible assessment of both human-authored and AI-generated

AD. By incorporating delivery-related dimensions in addition to traditional content-based measures, it supports a more comprehensive examination of quality in long-form video contexts.

Dataset Construction

The dataset used in this study consisted of ten publicly available videos selected to represent a range of genres, audiences, and quality levels. The goal was to assemble a diverse yet balanced set of materials that could effectively test the robustness of the evaluation framework across different descriptive challenges. The selection process followed three stages.

First, a large pool of candidate videos with volunteer-authored audio descriptions was collected from the YouDescribe [53] platform. Each video on the platform is rated on a 1–5 scale by viewers, providing a rough signal of description quality. Videos with ratings between 2 and 5 were chosen to capture variability across the quality spectrum.

Second, an accessibility consultant with professional experience in audio description prescreened the videos to ensure the inclusion of samples with distinct descriptive strengths and weaknesses. This step was used only for preliminary sorting rather than as part of the formal evaluation data, ensuring that the final selection retained natural variation while avoiding extreme outliers.

Third, ten videos were curated to cover three major categories: Entertainment, How-to & Style, and Education. The entertainment group included film trailers such as "*Star Wars:*

Entertainment

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

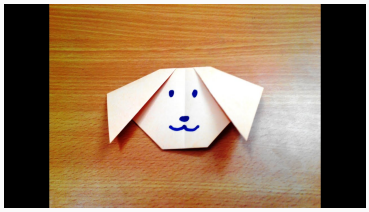
(b) Lady Bird | Official Trailer HD | A24

(c) Frozen Teaser (2013) - Disney Animated Movie

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

How-to & Style

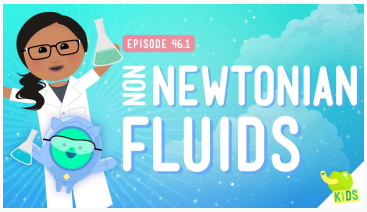
(e) 3 Ways to Make Homemade Pickles



(f) How to Make an Origami Dog Face

(g) Quick and Easy 5-Minute Makeup Tutorial

Education



(h) Non-Newtonian Fluids: Crash Course Kids

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

(j) jane goodall

Figure 6.2: Ten YouTube videos used in the evaluation, grouped by category: Entertainment (top), How-to & Style (middle), and Education (bottom). Subcaptions show the YouTube titles.

The Rise of Skywalker", "*Lady Bird*", "*Frozen*", and "*Elf*". The how-to group consisted of short instructional videos like "*3 Ways to Make Homemade Pickles*", "*How to Make an Origami Dog Face*", and "*Quick and Easy 5-Minute Makeup Tutorial*". The educational group contained informational videos such as "*Non-Newtonian Fluids – Crash Course Kids*",

"*Bald Eagle | Animals for Kids*", and "*Jane Goodall at Gombe*".

For each of the ten selected videos, 4 different AD versions were prepared, resulting in 40 total items (10 videos × 4 ADs) for evaluation. Each video included one human-authored version and three versions generated by VLMs.

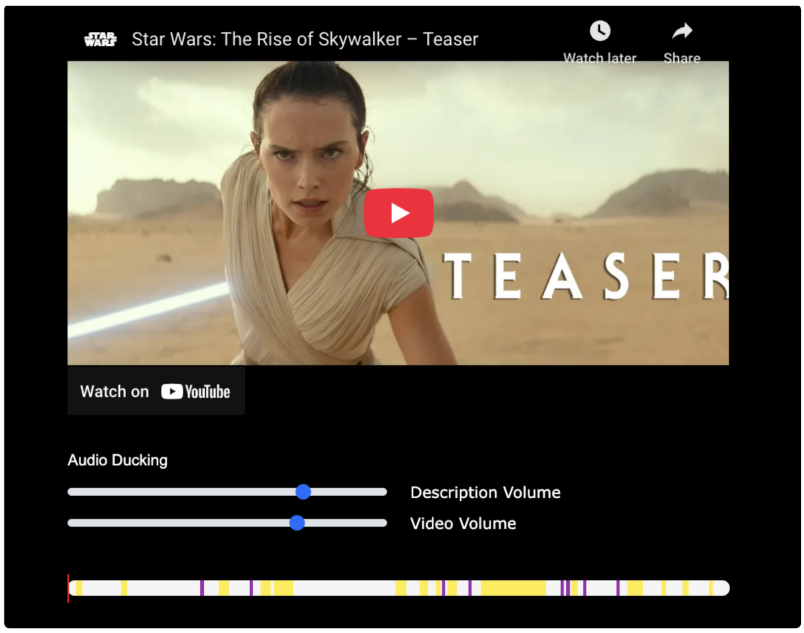
- **Volunteer-Authored Audio Description**: The human-authored versions were drawn directly from the YouDescribe platform. Each recording was automatically transcribed using Whisper, an open-source speech-to-text model with approximately 98 percent transcription accuracy. Minor transcription errors were manually corrected to maintain fidelity to the original volunteer narration. To eliminate potential bias caused by voice characteristics or recording quality, all descriptions were re-synthesized using a consistent synthetic voice through Google Cloud Text-to-Speech. Speech rate and intonation were adjusted to reflect the pacing and intent of the original performance.

- **VLM-generated Audio Description**: Three state-of-the-art VLMs, Qwen 2.5-VL, Gemini 1.5 Pro, and GPT-4o, were used to generate alternative descriptions. Each video was segmented into coherent scenes based on visual similarity and dialogue alignment, forming structured inputs for the models. Prompts were adapted from professional AD guidelines, emphasizing factual correctness, conciseness, neutrality, and synchronization with the video timeline while discouraging hallucination and redundancy. Once descriptions were generated, they were refined to reduce repetition and maintain narrative flow, then converted into synthetic speech using the same TTS

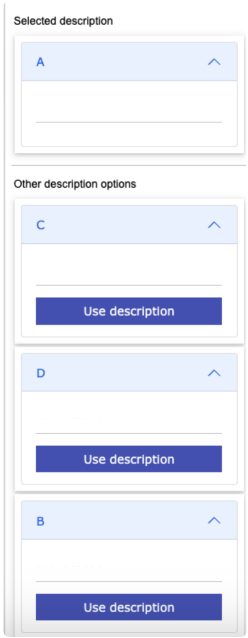
configuration applied to the human recordings. Each resulting audio track supported both inline and extended narration styles.

Rating Collection

Human Respondents



(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.

Figure 6.3: Viewing interface for synchronized playback of audio description with video.

All ratings from three human experts for the ground truth and the four additional respondents were collected through a custom-built evaluation interface developed for synchronized playback of video and AD. As shown in figure 6.3a, the interface displayed inline narration segments as yellow overlays on the video timeline and purple markers for extended narra-

tion that paused playback briefly to deliver additional descriptive content. Adjustable audio ducking controls allowed raters to balance narration volume against the original soundtrack, ensuring that both the content and timing of the descriptions could be clearly perceived. Furthermore, each video contained four available AD versions: one volunteer-authored and three generated by VLMs which are represented as A, B, C, and D (see figure 6.3b.

During the rating collection, 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.

Furthermore, each human respondent 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 shown in figure 6.3b. Human respondents 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.

Lastly, to establish the ground truth for analysis, majority/consensus rule was applied to the 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, the median of three expert ratings were collected to ensure 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 this case study.

VLM Respondents

```
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)
"""
```

Figure 6.4: A snippet of the prompt used to instruct the VLMs at applying the 7 dimensional assessment framework to evaluate ADs.

In addition to human participants, three VLMs, Qwen 2.5-VL, Gemini 1.5 Pro, and GPT-4o, were employed as automated raters. These models were selected for their complementary architectures and input capabilities, allowing comparison across different scales of multimodal reasoning. Each model was instructed to perform the same evaluation task as the human respondents using the seven-dimensional assessment framework described earlier.

Each VLM received a structured prompt containing the complete assessment framework, including the seven dimensions, their definitions, scoring criteria on a 1–5 ordinal scale, and illustrative examples (Figure 6.4, detailed in Appendix .7). The prompt specified that the model would be given two structured inputs:

1. the dialogue transcript extracted from the original video audio track, and
2. the JSON metadata describing each AD segment, including the narration text, start and end timestamps, delivery type (inline or extended), and description category (visual versus text-on-screen).

Two different role framings were used to examine how task framing influenced model behavior.

- In the **standard** framing (first version of system prompt): *"You are an expert Accessibility Consultant specializing in the quality assurance of audio description (AD) for video content."*
- In the **strict** framing (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."*

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 6.1: Eight VLM evaluation conditions combining model, input format, and role framing.

The stricter framing was introduced after preliminary testing revealed a bias toward high ratings (predominantly 4s and 5s). The addition of this constraint produced a broader distribution of scores and enhanced the discriminative power of the analysis.

Because each VLM differs in its ability to process audio and video inputs, four input conditions were developed, resulting in eight total evaluation configurations (three models $\times$ two framings $\times$ different video formats).

- Qwen 2.5-VL could not process full-length video due to GPU memory limitations. Each video was divided into 30-second segments, and the model evaluated the corresponding JSON data and video snippets. Segment-level ratings were averaged to produce video-level scores.
- GPT-4o shared similar token and frame-limit constraints. Videos were also segmented into 30-second intervals, and up to 30 frames per segment were supplied as visual input. Segment ratings were aggregated to obtain overall scores.
- Gemini 1.5 Pro accepted entire video uploads, allowing end-to-end evaluation of the full content in one session. In addition to this JSON + video input condition, a second variant was tested where the model received a screen-recorded playback video that included the synchronized audio, description, and timeline markers; the same interface used by human raters. This condition provided the closest approximation of the multimodal experience encountered by human participants.

Partial-Credit Scoring Construction

All collected ratings, those from the four additional human respondents and from the VLMs, were compared against the expert-established ground-truth ratings to construct the PCM. Consistent with the previous case studies, each response was converted into a three-level alignment score, reflecting how closely the rating matched the expert reference: 2 for exact agreement, 1 for adjacent agreement (one point difference in either direction), and 0 for distal agreement (a difference of two or more points). This recoded data then was applied to IRT Framework for each dimension.

6.2 Result and Discussion

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

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 6.2). 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 6.2). 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.

Person-Level Reliability and Fit

At the respondent level, the Mean Square (MNSQ) fit statistic [**wu2013properties**] (see Table .3 in Appendix .3) provides a quantitative check on the validity of the Item-Respondent map. The acceptable range is 0.75 to 1.33, with 1.0 representing the ideal value. Items with a fit statistic above 1.33 indicate excessive randomness. For example,

high-ability respondents performing poorly and low-ability respondents performing well. In contrast, values below 0.75 suggest underfit, meaning the item is overly predictable. One exception was Human4, whose fit statistics exceeded the acceptable range on Accurate, Appropriate and Equal dimensions. Thus, this should be considered when interpreting this person'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.

Wright Map Interpretation

As previously mentioned, there are total 7 dimension of the assessment framework. Because each dimension represents a distinct construct of audio description quality, the analyses were treated as independent models rather than components of a single composite scale which resulted in 7 distinct Wright Maps. Each Wright Map catures respondent performance and item difficulty specific to its respective dimension. Additionally, the items on the Wright Maps are ordered by their thresholds from lowest to highest to have better interpretability.

Accurate

Figure 6.5 presents the Wright Map for the *Accurate* dimension. Most respondents are distributed between logits -1 and $+1$, indicating a generally good level of agreement with the expert benchmark on this scale. The clustering of both human and VLM respondents within this central region suggests that the majority performed consistently and demonstrated com-

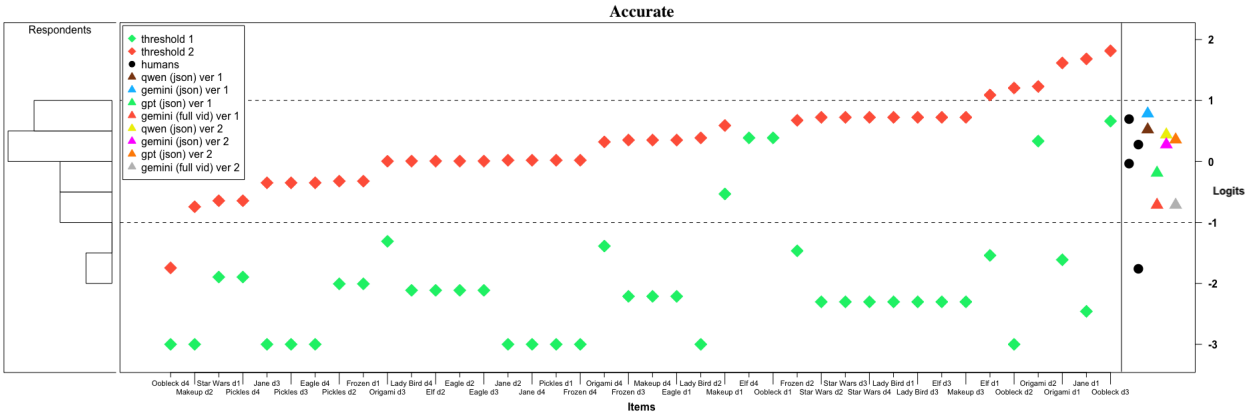


Figure 6.5: Wright Map for *Accurate* dimension showing most respondents clustered between -1 and $+1$ logits.

parable proficiency in judging factual correctness. The only clear outlier is Human4, located at the lower end of the scale. As shown in Appendix .3, this rater's fit statistic (MNSQ $=$ 2.3) exceeds the acceptable range, implying irregular or inconsistent rating behavior; thus, this individual's placement is not interpreted further. Among the model-based respondents, Gemini (JSON ver. 1) achieved the highest proficiency estimate, aligning exactly with the ground-truth ratings for roughly 34 out of 40 items. Several other VLMs also showed strong alignment, achieving more than half of the items in exact agreement. Overall, VLM respondents aligned closely with human raters, indicating that they evaluated description accuracy reliably and at a comparable level of proficiency.

Prioritized

The Wright Map for the *Prioritized* dimension (Figure 6.6) reveals limited interpretive value. The respondents are heavily clustered within a narrow logit range, offering minimal

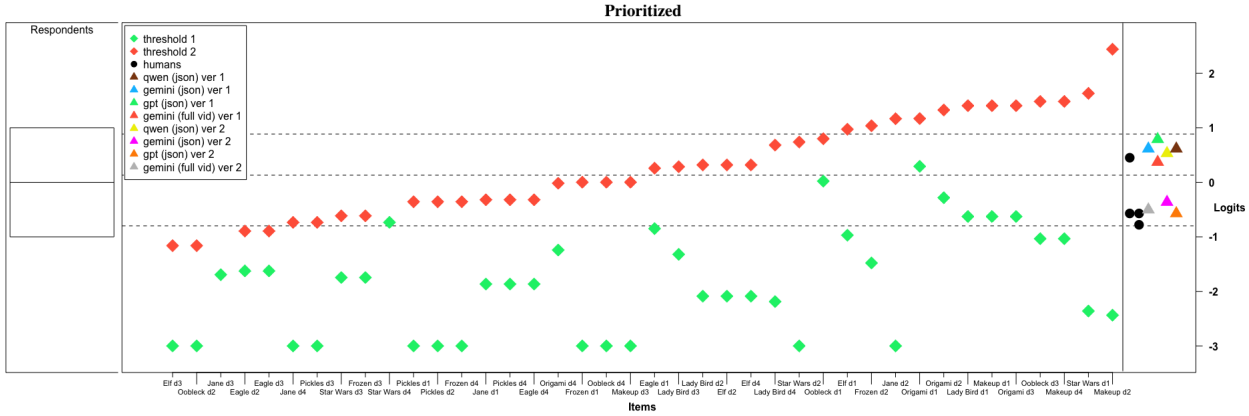


Figure 6.6: Wright Map for *Prioritized* dimension showing narrow region of respondents location with the cutpoints.

separation across ability levels. This lack of dispersion indicates that the scale provided little differentiation among raters, making it difficult to identify meaningful performance differences. The concentration of both thresholds and respondents suggests that this dimension may not effectively capture variation in evaluative ability. The respondent with the highest proficiency estimate (θ) is GPT (JSON ver. 1); however, even this configuration struggled with more than one quarter of the items, failing to achieve consistent exact agreement. Because the cutpoints are compressed and there is limited information above or below the main band of logits, this map yields low interpretability. The cause could stem from the scale design, item characteristics (i.e., the specific audio descriptions selected), or possibly rater variability. While the ground-truth ratings are less likely to be the source of this compression, the overall pattern suggests that the *Prioritized* dimension in its current form may require further refinement to produce more discriminative and interpretable results.

Appropriate

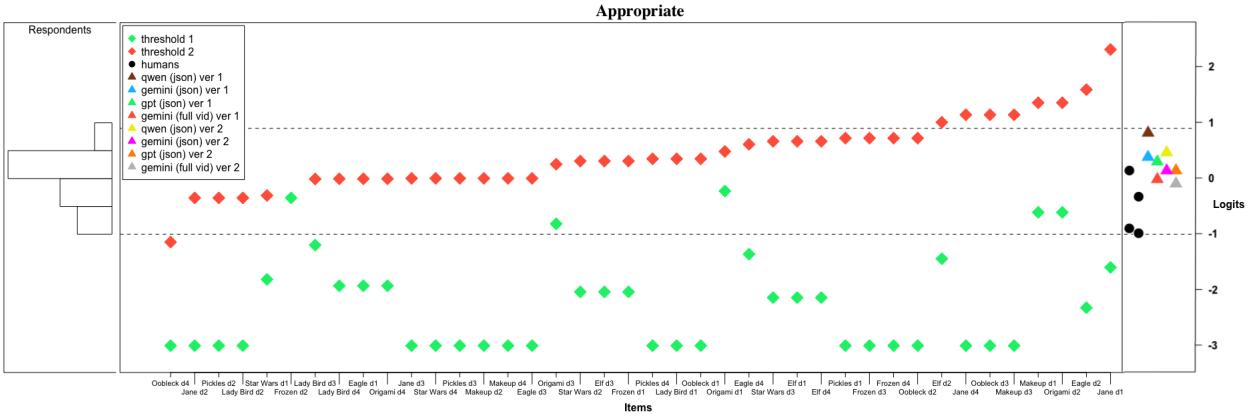


Figure 6.7: Wright Map for *Appropriate* dimension showing most of the respondents achieved the average or slightly above than the average on the AD rating tasks.

The Wright Map for the *Appropriate* dimension (Figure 6.7) initially appears moderately balanced, with respondents spread across the central logit range rather than being tightly clustered. However, upon closer inspection (see Appendix .4), this dimension exhibits the lowest number of well-fitting items, with many showing Item–Rest Correlations below 0.2. This pattern indicates that several items did not align consistently with the overall measurement scale. In practical terms, respondents who generally performed well sometimes received lower scores on specific items, while weaker respondents achieved unexpectedly high scores. Although Qwen (JSON ver. 1) showed slightly higher proficiency and achieved roughly 32 of 40 items in exact agreement, the irregular relationship between item difficulty and respondent ability reduces the interpretability of this dimension. While the Wright Map may look structurally acceptable, the underlying statistics suggest instability in how the *Appropriate* dimension captures rater performance. This inconsistency could stem from ambiguous item

phrasing or uneven difficulty calibration, indicating that this dimension may require further refinement before being treated as a reliable scale.

Consistent

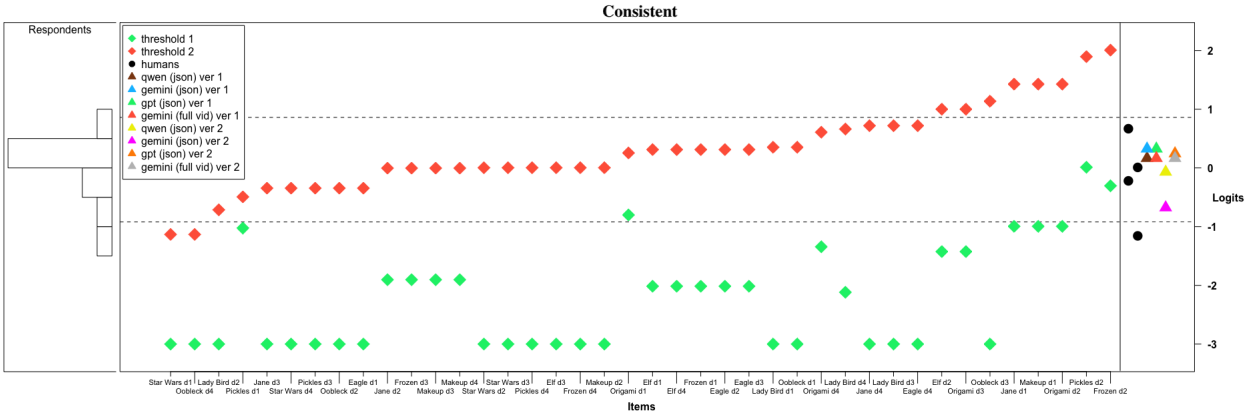


Figure 6.8: Wright Map for *Consistent* dimension showing most of the respondents achieved the average on the AD rating tasks.

The Wright Map for the Consistent dimension (Figure 6.8) displays a pattern similar to that observed in the Appropriate dimension. Most respondents are clustered near the 0 logit mark, showing minimal spread across the proficiency scale. Only one respondent, Human4, appears notably below this cluster, consistent with the earlier observation of irregular rating behavior. As shown in Appendix .4, only 17 out of 40 items in this dimension achieved an Item–Rest Correlation of 0.2 or higher. This relatively low proportion suggests limited internal consistency among items and weak discrimination power in separating raters by ability. Consequently, while the map shows general alignment among respondents, the low reliability of items makes it difficult to draw meaningful conclusions about the latent trait

measured by this dimension. The results imply that the Consistent dimension may not provide sufficient information to differentiate evaluator performance effectively.

Equal

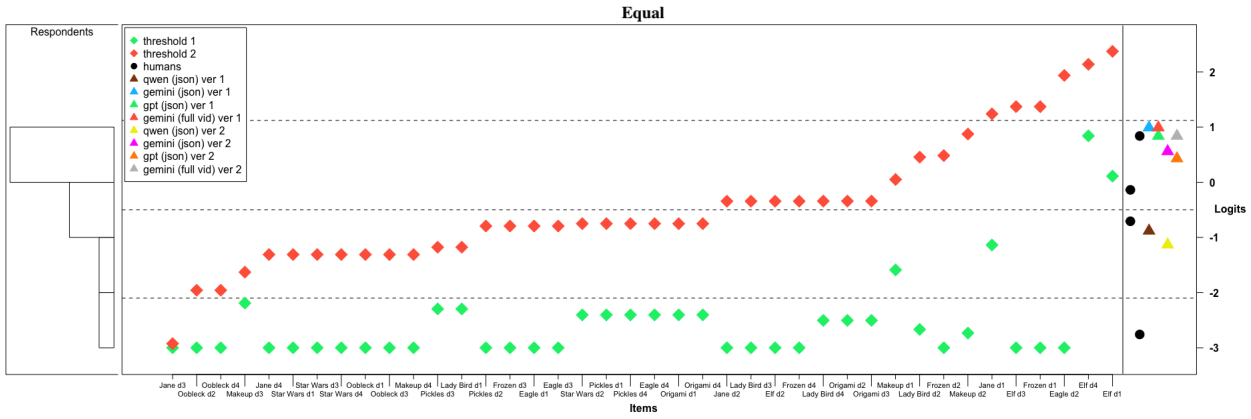


Figure 6.9: Wrightmap for *Equal* dimension which shows huge variance of respondents location. It is informative and has rich interpretability to analyze.

The Wright Map for the Equal dimension (Figure 6.9) demonstrates the widest variance in respondent locations across the logit scale, indicating that this dimension provided the richest and most informative measurement among all seven. The highest proficiency estimates (θ=0.99126) were observed for Gemini (JSON ver. 1) and Gemini (Full Video ver. 1), each achieving exact agreement on roughly 34 of 40 items. 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. Overall, the Equal dimension produced the most interpretable and reliable Wright Map, capturing meaningful variation in evaluative ability across both human and model raters.

Strategy

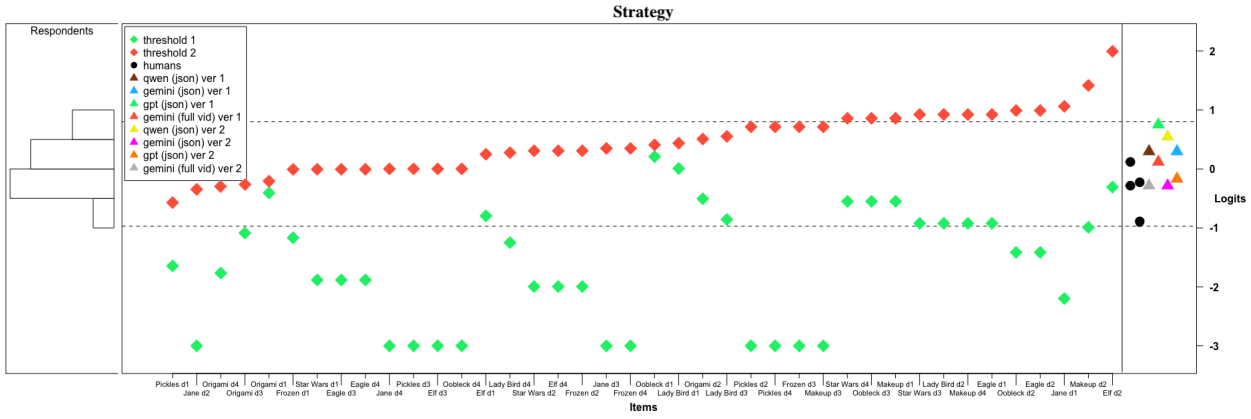


Figure 6.10: Wright Map for *Strategy* dimension showing most of the respondents achieved the average or slightly above than the average on the AD rating tasks.

The Wright Map for the *Strategy* dimension (Figure 6.10) shows a moderate clustering of respondents around the average logit region, similar in overall structure to the *Appropriate* dimension. Item reliability is likewise limited, with the same number of items reaching an

Item–Rest Correlation of 0.2 or higher, suggesting that this dimension also provides only modest differentiation among raters. The highest proficiency estimate (θ=0.75135) was recorded for GPT (JSON ver. 1), which achieved exact agreement on slightly fewer than three-quarters of the items. Despite its relatively strong performance, the narrow distribution of respondents and limited item spread indicate that this dimension does not clearly separate high and low performing evaluators.

Timing

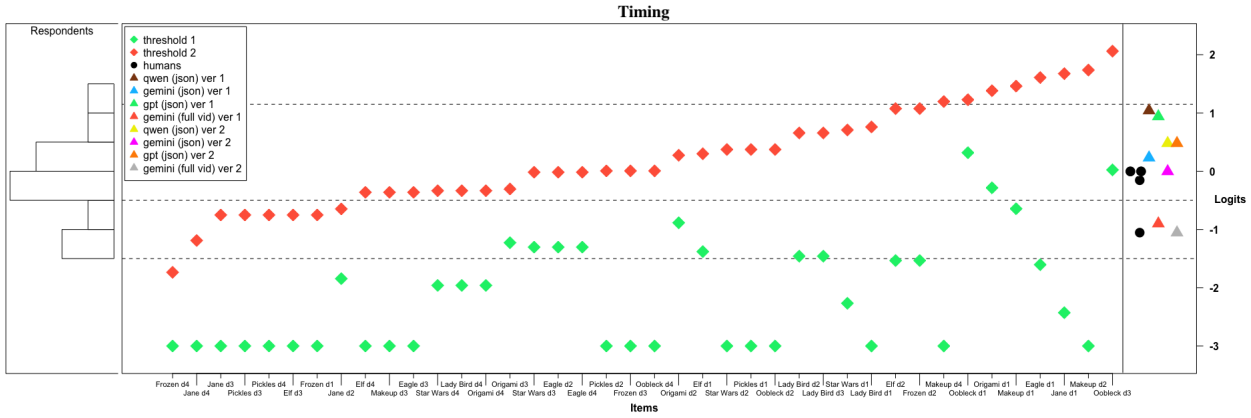


Figure 6.11: Wright Map for *Timing* dimension showing wide range of respondent location which indicates that this is informative

The Wright Map for the Timing dimension (Figure 6.11) shows a relatively wide spread of respondent locations across the logit scale, indicating noticeable variation in evaluator proficiency. Based on the cut points, respondents appear to form two general groups: one cluster performing above the average range and another slightly below it. Although there are no respondents located in the lower logit region, the existing distribution still provides suffi-

cient variance for meaningful interpretation. Among all raters, Qwen (JSON ver. 1) achieved the highest proficiency estimate (θ=1.04167), reflecting strong alignment with the ground-truth ratings. Overall, this dimension demonstrates clearer separation between respondents compared to several other dimensions and provides a reasonably interpretable measure of how consistently evaluators, both human and VLM, assessed the temporal placement and synchronization of the audio descriptions.

6.3 Limitation and Future Work

While this study provided a structured examination of how VLMs and human raters evaluate audio descriptions using IRT, several limitations should be acknowledged.

First, the sample size was modest in both the number of items and respondents. With only ten videos and forty audio-description versions evaluated by a small cohort of human and VLM raters, the statistical power was limited. This constraint prevented more detailed comparisons across genres or subgroups of evaluators. Future work should expand both item diversity and respondent representation, incorporating a wider range of video types, varying lengths, and additional domains of AD. On the respondent side, including novice describers, professional describers, and BLV users would provide a more comprehensive understanding of how different groups assess AD quality. Likewise, extending this approach to newer and more varied VLMs will help clarify model-specific behaviors and progress over time.

Second, the analysis revealed variation in item fit quality, suggesting that some descrip-

tions functioned more effectively than others in distinguishing evaluator ability. Items with low fit or weak correlation values likely introduced additional noise to the model, making certain dimensions harder to interpret. Future iterations should investigate why certain AD items align poorly with rater performance and consider removing or recalibrating such items before reconstructing the Rasch model to improve precision and reliability.

Third, while the seven dimensional assessment framework successfully integrates both content and formatting aspects of AD evaluation, its psychometric properties have yet to be fully validated. Establishing strong reliability, construct validity, and criterion validity will require larger datasets than those used in this study. Future work should focus on collecting more extensive rating data to enable robust validation through methods such as factor analysis or multi-dimensional IRT modeling.

Despite these limitations, this study provides valuable implications for hybrid human–AI evaluation workflows. Rather than replacing human expertise, VLMs can serve as scalable first-pass evaluators, efficiently identifying potential issues before expert review. Dimensions such as timing and delivery strategy, often overlooked in previous frameworks which proved critical for understanding real-world accessibility and should remain central in future models. As VLMs continue to evolve, developing evaluation workflows that combine their efficiency with the diagnostic insight of human raters will be essential for creating sustainable, high-quality accessibility assessment systems.

Chapter 7

Conclusion

This thesis presented a series of empirical case studies applying IRT to evaluate the performance of VLM as raters in multimodal accessibility contexts. Across image, comic, and video modalities, psychometric modeling was shown to provide a principled framework for comparing AI and human evaluation behavior on a shared latent scale of proficiency. By treating models as respondents and their outputs as structured responses to defined items, IRT enabled a deeper examination of evaluative consistency, sensitivity, and bias beyond surface-level accuracy metrics. This methodological approach underscores the potential of IRT as a rigorous quantitative lens for assessing model reliability and interpretive behavior, revealing how VLM decision patterns approximate, diverge from, or even surpass human evaluative tendencies across modalities and task complexities.

Across the three case studies, the results collectively indicate that VLMs are becoming increasingly capable of functioning as consistent and reliable evaluators across diverse

multimodal contexts. Although continued refinement in dataset scale, rater diversity, and dimension-specific validation is necessary, the findings point to the growing potential of VLMs as scalable and psychometrically grounded raters for visual and accessibility-oriented evaluation tasks.

The integration of IRT into VLM assessment marks a shift from accuracy-based benchmarking toward measurement-based understanding of model behavior. By positioning human and machine raters on a common latent scale, IRT enables interpretable analysis of how evaluators differ not only in outcomes but in consistency and sensitivity to task difficulty. This perspective provides a principled foundation for assessing how far AI raters can emulate human evaluative judgment while also exposing the internal validity of the scales themselves.

Yet, human oversight remains essential for interpretability, ethical review, and cultural nuance. Rather than replacing humans, future evaluation systems should adopt hybrid workflows in which VLMs provide scalable quantitative ratings and humans contribute qualitative reasoning and normative grounding. Such complementarity ensures that evaluation remains both efficient and accountable, leveraging the computational reach of VLMs while retaining the contextual judgment that defines human expertise. As multimodal AI systems continue to evolve, integrating human-centered measurement theory will be crucial to ensuring that the expansion of automated evaluation advances not only performance but also fairness, accountability, and inclusivity in accessible technology.

Bibliography

[1] 3Play Media. *Audio Description (AD) Guidelines*. 2020. URL: https://www.3playmedia.com/popular-topics/audio-description/.

[2] Google AI. *Gemini 1.5 Flash*. Accessed: 2025-01-22. 2024. URL: https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-flash.

[3] Alibaba Cloud Research Institute. *Qwen2.5-VL: A Large Multimodal Model for Image, Video, and Text Understanding*. https://huggingface.co/Qwen/Qwen2.5-VL. Open-source Vision-Language Model capable of multimodal reasoning over image and video inputs. Released September 2024. Sept. 2024.

[4] Peter Anderson et al. "Spice: Semantic propositional image caption evaluation". In: *European conference on computer vision*. Springer. 2016, pp. 382–398.

[5] David Andrich. "Rasch rating-scale model". In: *Handbook of item response theory*. Chapman and Hall/CRC, 2016, pp. 75–94.

[6] Stanislaw Antol et al. "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[7] Ge Bai et al. "Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues". In: *arXiv preprint arXiv:2402.14762* (2024).

[8] Daniel Bergin and Brett Oppegaard. "Automating Media Accessibility: An Approach for Analyzing Audio Description Across Generative Artificial Intelligence Algorithms". In: *Technical Communication Quarterly* 34.2 (2025), pp. 169–184. ISSN: 15427625. DOI: 10.1080/10572252.2024.2372771.

[9] Carmen J Branje and Deborah I Fels Structured. *LiveDescribe: Can Amateur Describers Create High-Quality Audio Description?* Tech. rep.

[10] Lin Chen et al. "Sharegpt4video: Improving video understanding and generation with better captions". In: *arXiv preprint arXiv:2406.04325* (2024).

[11] Wei-Lin Chiang et al. "Chatbot arena: An open platform for evaluating llms by human preference". In: *Forty-first International Conference on Machine Learning*. 2024.

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

[13] Neil Cohn. "A visual lexicon". In: *Public Journal of Semiotics* 1.1 (2007), pp. 35–56.

[14] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[15] Michael Denkowski and Alon Lavie. "Meteor universal: Language specific translation evaluation for any target language". In: *Proceedings of the ninth workshop on statistical machine translation*. 2014, pp. 376–380.

[16] *Described and Captioned Media Program (DCMP)*. 2024. URL: `https://dcmp.org/learn/descriptionkey`.

[17] Nazaret Fresno. "Closed captioning quality in the information society: the case of the American newscasts reshown online". In: *Universal Access in the Information Society* 20.4 (2021), pp. 647–660.

[18] *Fundamentals of item response theory*. URL: `https://psycnet.apa.org/record/1991-98425-000`.

[19] Google DeepMind. *Gemini 1.5 Pro: A Multimodal Large Language Model with Long-Context and Video Understanding Capabilities*. `https://deepmind.google/technologies/gemini/`. Closed-source multimodal model supporting text, image, audio, and video processing. Released February 2024. Feb. 2024.

[20] Micah Hodosh, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics". In: *Journal of Artificial Intelligence Research* 47 (2013), pp. 853–899.

[21] Maurice G Kendall. "A new measure of rank correlation". In: *Biometrika* 30.1/2 (1938), pp. 81–93.

[22] Chaoyu Li et al. "VideoA11y: Method and Dataset for Accessible Video Description". In: *Conference on Human Factors in Computing Systems - Proceedings* (Apr. 2025). DOI: 10.1145/3706598.3714096/SUPPL{_}FILE/PN2974-TALK-VIDEO.MP4. URL: /doi/pdf/10.1145/3706598.3714096?download=true.

[23] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81.

[24] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[25] Jing Liu et al. "VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47.2 (2025), pp. 708–724. ISSN: 19393539. DOI: 10.1109/TPAMI.2024.3479776.

[26] Geoff N Masters. "A Rasch model for partial credit scoring". In: *Psychometrika* 47.2 (1982), pp. 149–174.

[27] Geofferey N. Masters and Benjamin D. Wright. "The Partial Credit Model". In: *Handbook of Modern Item Response Theory* (1997), pp. 101–121. DOI: 10.1007/978-1-4757-2691-6{_}6. URL: https://link.springer.com/chapter/10.1007/978-1-4757-2691-6_6.

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

[29] Inc. Meta Platforms. *Llama-3.2-11B-Vision-Instruct*. Accessed: 2025-01-22. 2024. URL: https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.

[30] Microsoft. *Phi-3.5 Vision Instruct*. Accessed: 2025-01-22. 2024. URL: https://huggingface.co/microsoft/Phi-3.5-vision-instruct.

[31] Cosmin Munteanu et al. "Situational ethics: Re-thinking approaches to formal ethics requirements for human-computer interaction". In: *Conference on Human Factors in Computing Systems - Proceedings* 2015-April (Apr. 2015), pp. 105–114. DOI: 10.1145/2702123.2702481;CTYPE:STRING:BOOK. URL: /doi/pdf/10.1145/2702123.2702481?download=true.

[32] Sawako Nakajima and Kazutaka Mitobe. "Professional and novice audio describers: quality assessments and audio interactions". In: *Journal of Specialised Translation* 42 (July 2024), pp. 64–83. ISSN: 1740357X. DOI: 10.26034/CM.JOSTRANS.2024.5980. URL: https://www.researchgate.net/publication/382718827_Professional_and_novice_audio_describers_quality_assessments_and_audio_interactions.

[33] Lothar D Narins et al. "Validated image caption rating dataset". In: *Advances in Neural Information Processing Systems* 36 (2024).

[34] OpenAI. *GPT-4o: A Unified Multimodal Model for Text, Image, Audio, and Video Understanding*. https://openai.com/research/gpt-4o. Closed-source Vision-Language Model supporting real-time multimodal reasoning. Released May 2024. May 2024.

[35] OpenAI. *OpenAI Models Documentation*. Accessed: 2025-01-22. 2025. URL: https://platform.openai.com/docs/models.

[36] Jaclyn Packer, Katie Vizenor, and Joshua A Miele. "An overview of video description: history, benefits, and guidelines". In: *Journal of Visual Impairment & Blindness* 109.2 (2015), pp. 83–93.

[37] Kishore Papineni et al. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[38] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. "Rescribe: Authoring and automatically editing audio descriptions". In: *UIST 2020 - Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Oct. 2020), pp. 747–759. DOI: 10.1145/3379337.3415864/SUPPL{_}FILE/3379337.3415864.MP4. URL: https://dl.acm.org/doi/10.1145/3379337.3415864.

[39] Hugging Face TB Research. *SmolVLM-Instruct*. Accessed: 2025-01-22. 2024. URL: https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct.

[40] Anna Rohrbach et al. "Movie Description". In: *International Journal of Computer Vision* 123.1 (May 2017), pp. 94–120. ISSN: 15731405. DOI: 10.1007/S11263-016-0987-1. URL: https://dl.acm.org/doi/10.1007/s11263-016-0987-1.

[41] FUMIKO Samejima. "Graded response model of the latent trait theory and tailored testing". In: *Proceedings of the first conference on computerized adaptive testing*. US Government Printing Office Washington DC. 1976, pp. 5–17.

[42] Martin Schmettow and Wolfgang Vietze. "Introducing Item Response Theory for measuring usability inspection processes". In: *Conference on Human Factors in Computing Systems - Proceedings* (2008), pp. 893–902. DOI: 10.1145/1357054.1357196/SUPPL{_}FILE/P893-AUDIOCHANNEL0.MP3. URL: /doi/pdf/10.1145/1357054.1357196?download=true.

[43] Andrew Taylor Scott et al. "Improved image caption rating–datasets, game, and model". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–7.

[44] Charles Spearman. "The proof and measurement of association between two things". In: *The American Journal of Psychology* 15.1 (1904), pp. 72–101. DOI: 10.2307/1412159.

[45] Qwen Team. *Qwen2-VL-7B-Instruct*. Accessed: 2025-01-22. 2024. URL: https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct.

[46] Tess Van Daele et al. "Making Short-Form Videos Accessible with Hierarchical Video Summaries". In: *Conference on Human Factors in Computing Systems - Proceedings* 1 (Feb. 2024), p. 17. DOI: 10.1145/3613904.3642839. URL: http://arxiv.org/abs/2402.10382%20http://dx.doi.org/10.1145/3613904.3642839.

[47] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

[48] Xin Wang et al. "VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research". In: *Proceedings of the IEEE International Conference on Computer Vision* 2019-October (Apr. 2019), pp. 4580–4590. ISSN: 15505499. DOI: 10.1109/ICCV.2019.00468. URL: https://arxiv.org/pdf/1904.03493.

[49] Yujia Wang and Wei Liang. "Toward automatic audio description generation for accessible videos". In: *Conference on Human Factors in Computing Systems - Proceedings* (May 2021). DOI: 10.1145/3411764.3445347/SUPPL{_}FILE/3411764.3445347{_}VIDEOPREVIEW.MP4. URL: https://dl.acm.org/doi/10.1145/3411764.3445347.

[50] Margaret A. Webb and June P. Tangney. "Too Good to Be True: Bots and Bad Data From Mechanical Turk". In: *Perspectives on Psychological Science* 19.6 (Nov. 2024), pp. 887–890. ISSN: 17456924. DOI: 10.1177/17456916221120027/ASSET/F0D2D167-57AC-453B-A3E6-16E5C2B3C2E3/ASSETS/IMAGES/LARGE/10.1177{_}17456916221120027-FIG1.JPG. URL: https://journals.sagepub.com/doi/10.1177/17456916221120027.

[51] Jun Xu et al. "Msr-vtt: A large video description dataset for bridging video and language". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5288–5296.

[52] Xiaojun Ye et al. *MMAD: Multi-modal Movie Audio Description*. Tech. rep. 2024, p. 11415. URL: https://github.com/Daria8976/MMAD..

[53] YouDescribe. *YouDescribe. Accessed Date 2025-03-08. https://www.youdescribe.org/.*

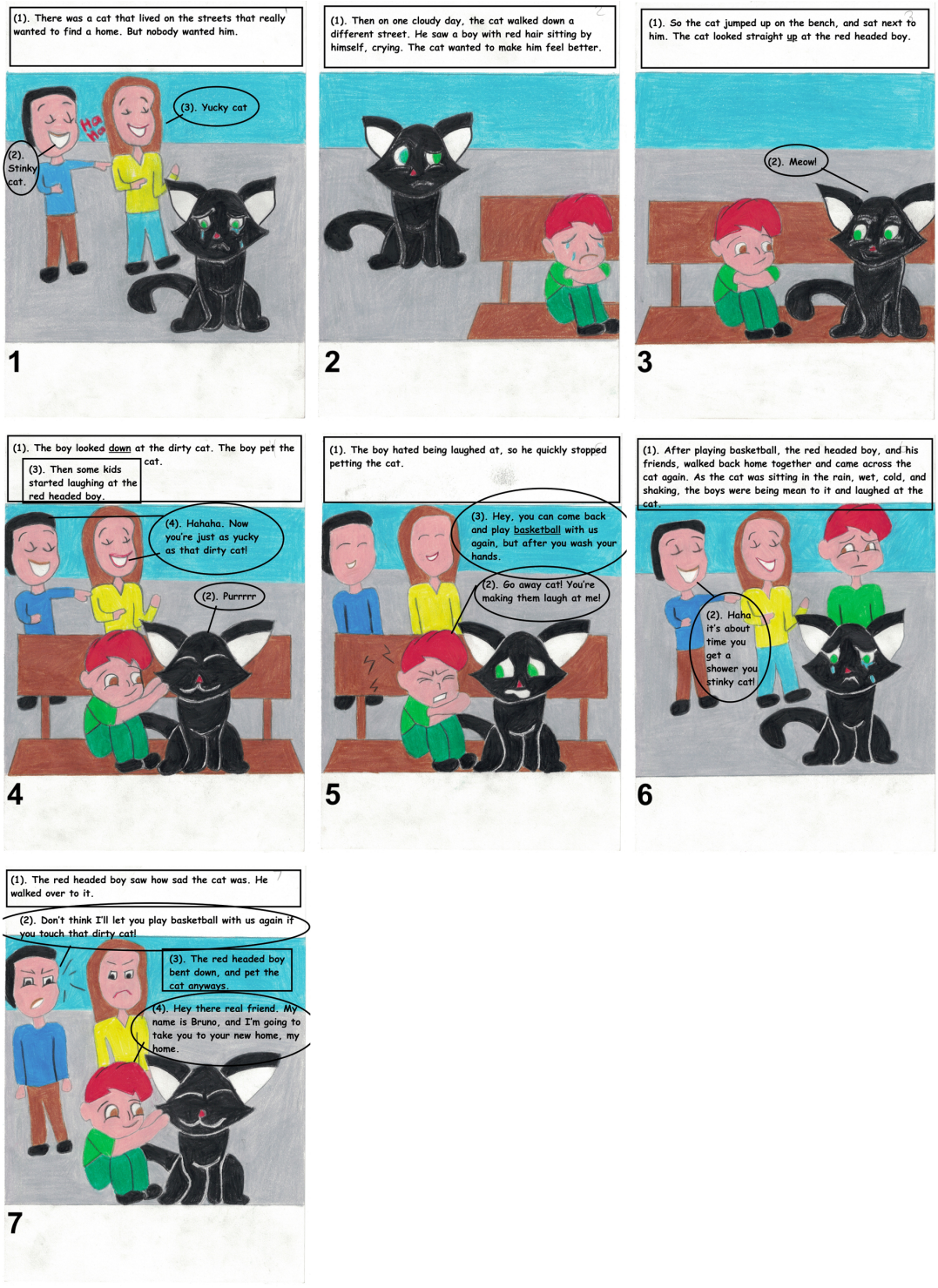
[54] Luowei Zhou, Chenliang Xu, and Jason J. Corso. "Towards Automatic Learning of Procedures from Web Instructional Videos". In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (Mar. 2017), pp. 7590–7598. ISSN: 2159-5399. DOI: 10.1609/aaai.v32i1.12342. URL: https://arxiv.org/pdf/1703.09788.

.1 Appendix A: 25 items on Case Study 1

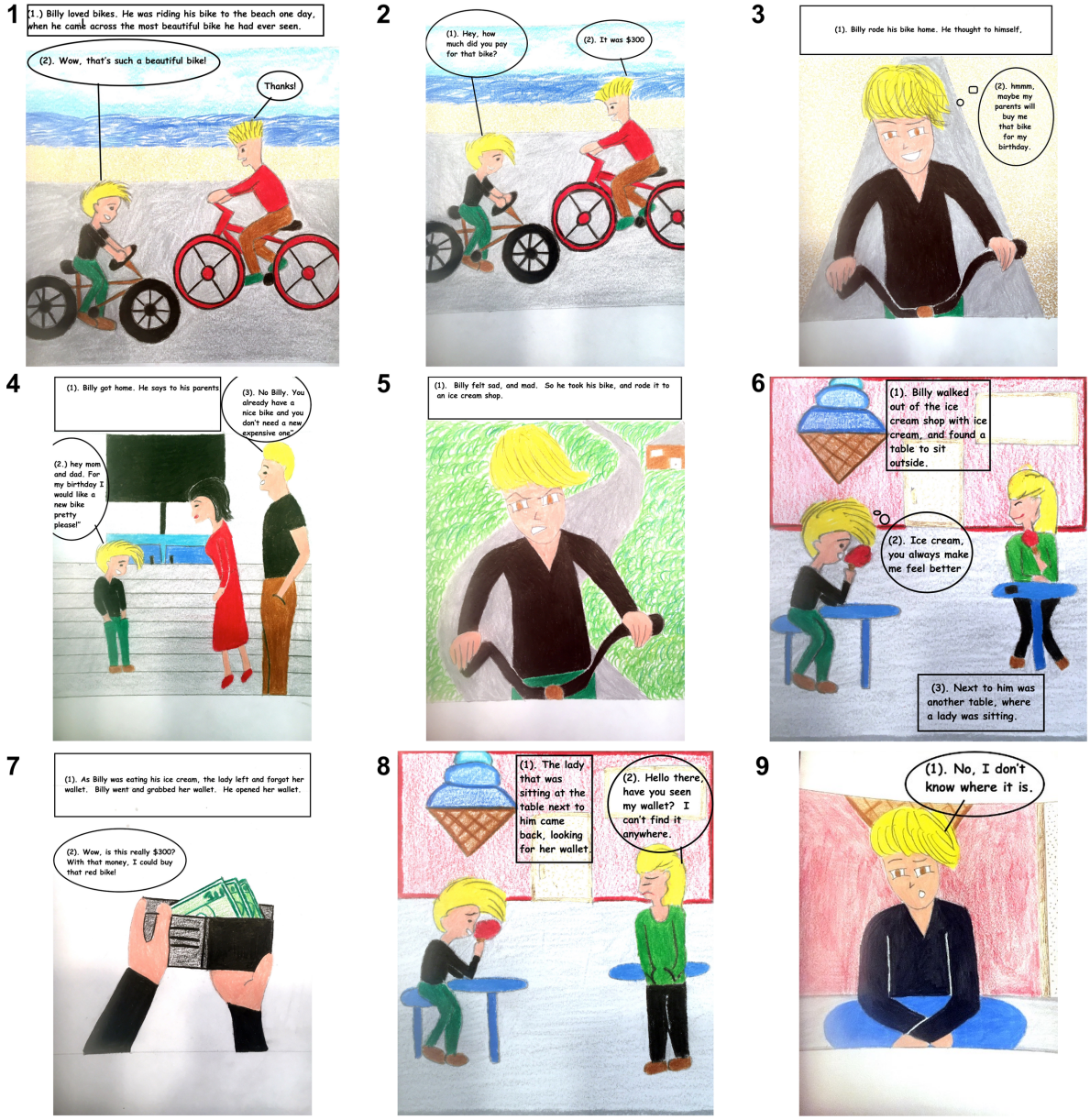
Image	Captions with 1 - 5 ratings
	**1.** Many white puppies are eating food near several brown roosters. **2.** A man with sunglasses is in a car. **3.** A person wearing red is in a car. **4.** A boy wearing a red shirt and socks is in a car. **5.** A boy wearing a soccer uniform holds muddy cleats and sits in a van with an open door.
	**1.** A young man jumping a back flip off of a concrete wall. **2.** A crowd of people at an outdoor event. **3.** Three people under an umbrella. **4.** Three people sit at a table under an umbrella. **5.** Three people sit at a table under an umbrella outside a cafe.
	**1.** The young football player is trying to avoid being tackled. **2.** A woman holding a young girl playing with bubbles at a picnic. **3.** A woman holding a young girl sit next to a basket. **4.** A woman holding a young girl and a woman wearing a red dress sit next to a basket of bread in a grassy field. **5.** Two women and a young girl wear costumes at a fair and crouch next to a wheelbarrow of bread.
	**1.** Team members being lifted up high to catch a flying ball. **2.** A man in grey on a white hill overlooking the ocean. **3.** A man sitting in snow. **4.** People walk in snow on a mountain with a pink sky. **5.** Snowboarders and skiers watch a sunset from a snow-covered mountain.
	**1.** A dog is running along the beach beside the ocean. **2.** A little girl kicks into the air. **3.** A girl jumps on the sand. **4.** A little girl runs on the wet sand near the ocean. **5.** A girl jumps toward the ocean waves on a sunny beach.

Table .1: 25 image-caption pairs selected by experts used in IRT analysis. Each Image contains 5 captions that correspond to 1-5 ratings.

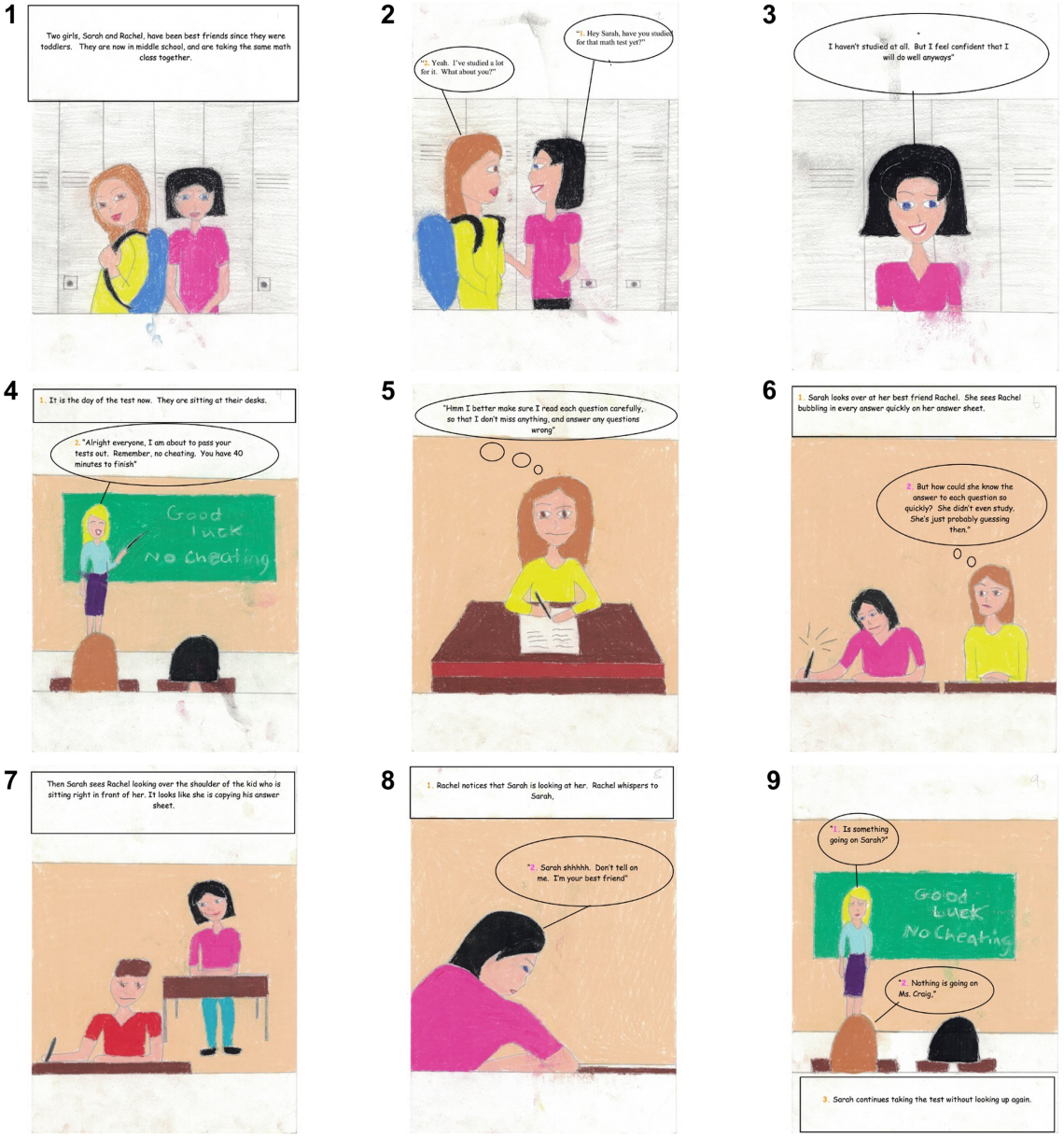
.2 Appendix B: Comic "Friendship"



.3 Appendix C: Comic "Stealing"



.4 Appendix D: Comic "Lying"



.5 Appendix E: VLMs Prompt on Case Study 2

```
System_prompt = """ You are a middle school or high school
    student. Please answer the following questions based
    on the given panels in 1 sentence.

<Comic Panels>

Question: <Motivationl Inference Question (Why)>
Assistant: Response 1
"""
```

↓

```
System_prompt = previous_systemp_prompt + """
Question: <Meta Reasoning (follow-up) Question: (What)>
Assistant: Response 2
"""
```

↓

```
System_prompt = previous_systemp_prompt + """
Question: <Evaluative Inference Question (What Lesson)>
Assistant: Response 3
"""
```

↓

```
System_prompt = previous_systemp_prompt + """
Question: <Meta Reasoning (follow-up) Question: (What)>
Assistant: Response 4
"""
```

Figure .1: The prompt was used as a chain where after 1st question was asked and retrieved the response, the prompt was stacked with the 2nd question and response and so on. This is to give all the previous context.

.6 APpendix F: Person Fit Statistic on Case Study 2

VLM Respondent	**Comic + Text**	**Text-Only**
GPT-4-Turbo (2024-04-09)	0.54884	0.72959
GPT-4o-mini	0.89652	0.95099
GPT-4o (2024-11-20)	0.84126	0.96616
GPT-4o (2024-08-06)	0.93006	0.84837
GPT-4o (2024-05-13)	0.50505	0.47726
Phi-3.5-Vision-Instruct	0.88509	0.74400
Qwen2-VL-7B-Instruct	0.79545	0.74063
Gemini 1.5 Flash	1.30588	0.84190
SmolVLM-Instruct*	0.63817	N/A

Table .2: 9 VLM respondents fit statistics where the ideal range is between 0.75 and 1.33 to be well-fit.

.7 Appendix G: VLM Respondents Prompt on AD Task

```
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."
"""
}}
```

.8 Appendix H: AD Person 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.

.9 Appendix I: AD 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 the internal consistency of person estimates, and should be > 0.90 to be a dependable predictor of person-ability. Well-fit items are defined as those with Item-Rest Cor. ≥ 0.20.