See both ways: A bidirectional evaluation of Multimodal Language Models and Human Spontaneous Speech for Image Captioning

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have achieved notable success in image captioning, yet systematic comparisons with human-generated references remain underexplored. In this work, we present a novel study on understanding the alignment between captions generated by multimodal models and spontaneous human speech captions. To this end, we introduce a human–machine bidirectional evaluation framework, which does not assume a "ground-truth". This evaluation is performed by comparing human audio-based captions of images with model generated captions from various MLLMs. Our detailed analysis reveals that, (i) humans are more selective than models in image captioning, rather than providing a comprehensive summary, (ii) scores with human reference and model targets are significantly higher than those computed with model reference and human targets, and (iii) images from specific categories like "nature" and "education" evoke more human imagination during the description task, compared to other categories. Together, these findings reveal a clear divergence in human vs. model captioning that can pave the way for human-aligned MLLM designs.

1 Introduction

Image captioning, the task of generating a verbal depiction of a visual presentation, is one of the most fundamental human skills. The spontaneous vocal description of images evokes visual attention, linguistic proficiency and memory, and is used in neuro-psychological assessments of cognitive skills (For example, the "Cookie Theft" picture description from the Boston-Diagnostic-Aphasia-Examination (Giles et al., 1996)). In natural image captioning, early work by (Griffin and Bock, 2000) suggested that eye movements during picture description predict the order of mention. (Huettig et al., 2011) explored the use of verbal descriptions of visual objects as a tool to probe planning processes. In non-English settings, a recent study by (Takmaz et al., 2024) on Dutch image description, attempted to relate the properties of an image and the human behavior of image description to quantify the visuo-linguistic complexity. Separately, (He et al., 2019) found that human attention, measured via eye-fixations, differs from regular viewing of images.

In machine learning, the image captioning task is also considered one of the challenging tasks that require visual and contextual understanding, saliency and relationship cognizance, and multi-modal alignment. Early approaches were template-based and more rigid in the caption generation (Farhadi et al., 2010). This was advanced by representation based approaches, like those investigated by (Kulkarni et al., 2011) and (Yang et al., 2011). As deep learning and sequence modeling architectures became popular, neural encoder-decoder models (for example, (Vinyals et al., 2015) and (Xu et al., 2015)) showed substantial improvements in free-form caption generation for natural images. Transformer models, like the proposal by (Anderson et al., 2018), further advanced this progress through attention based modeling. More recently, multi-modal models (Li et al., 2023a) and large language models (OpenAI, 2024) have become de facto image-captioning systems, owing largely to their improvements in understanding, reasoning and generation capabilities.

The evaluation of image captioning systems has similarly evolved, progressing from early text-based metrics such as BLEU (Papineni et al., 2002) and consensus-based approaches (Vedantam et al., 2015), to semantic similarity metrics (Anderson et al., 2016) and embedding-based methods (Hessel et al., 2021b). More recently, large language models have enabled caption evaluation in more human-aligned settings (Ye et al., 2025).

In this paper, we undertake a study on comparing human and model generated captions on natural,

culturally ingrained, and regionally relevant images. The dataset of images and the human captions are derived from the Vaani dataset (VAANI, 2025) and Places audio caption dataset (Harwath et al., 2016), where human participants provide audio captions of images in Hindi language. The proposed large-scale analysis in this paper, (10k image-transcript pairs and with 3288 human participants), involves the premise where both humans and models provide captions without a "ground-truth" setting. Hence, the study is singularly unique in various ways compared to previous studies.

The key contributions of this paper are:

1. We present a first-of-its-kind comprehensive benchmarking and anlysis of multimodal large language models on spontaneous, free-flowing, image-prompted speech, leveraging the Vaani dataset (VAANI, 2025), first of its kind for Indo-geographic and linguistic diversity. We also extend the analysis to Places audio caption dataset (Harwath et al., 2016).

2. We introduce a bidirectional evaluation framework based on two complementary metrics: the Human-as-Reference (HAR) score, which measures how closely model captions align with human references, and the Model-as-Reference (MAR) score, which evaluates the reverse alignment.

3. Using these scores, we conduct a detailed model-wise analysis of both open-source systems (GEMMA-3-12B and LLAMA-4-SCOUT-17B-16E-INSTRUCT) and closed-source systems (GEMINI 2.5 PRO and GPT-4O). Our study provides novel insights into differences in coverage, selectivity and hallucination profiles across models and datasets.

4. We perform a category-wise evaluation of human transcripts, examining how the image-caption quality varies across broad categories. This analysis highlights the differences in how humans and models prioritize visual content.

2 Related Work

Image Captioning Evaluation: Traditional approaches evaluating image captioning primarily rely on reference-based methods. BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and METEOR (Denkowski and Lavie, 2014) emphasize lexical overlap, thereby limiting their ability to capture deeper semantic attributes. Subsequent efforts, such as SPICE (Anderson et al., 2016), CLIP (Radford et al., 2021) and BLIP-2 (Li et al., 2023b) have improved the quality of the metrics. Examples include CLIPScore (Hessel et al., 2021a), PACScore (Sarto et al., 2023), and BLIP2Score (Zeng et al., 2024).

The integration of MLLMs into captioning evaluation has enabled a shift from reference-dependent metrics that ignore image content, as exemplified by CLAIR (Chan et al., 2023), to reference-free, image-grounded approaches that provide explainable scores, such as FLEUR (Lee et al., 2024). However, these explanations often lack standardized structure and remain largely unverified. To address this, EXPERT (Kim et al., 2025) proposes a framework for generating structured explanations evaluated along fluency, relevance, and descriptiveness dimensions. More recently, DCScore (Ye et al., 2025) explores concept-level quality assessment using LLMs. Despite these advances, a critical limitation in vision–language research persists: the linguistic and cultural bias inherent in canonical datasets, driven by an over-reliance on Anglocentric data sources (Liu et al., 2021).

Multimodal Large Language Models: Vision-Language Models (VLMs) (Dai et al., 2023; Li et al., 2023b; Jia et al., 2021; Yu et al., 2022) have substantially advanced image captioning tasks by developing large-scale, pre-trained highly capable Multimodal Large Language Models (MLLMs) (Bai et al., 2023b; Gao et al., 2023; Achiam et al., 2023; Chen et al., 2024a; Comanici et al., 2025). Models such as LLaVa (Liu et al., 2023b,a, 2024), Qwen-VL (Bai et al., 2023a), Intern-vl (Chen et al., 2024b) have accelerated the development of these general-purpose models.

Nevertheless, despite these advanced capabilities, MLLMs are susceptible to significant failure modes, including incomplete descriptions and object hallucinations. A major reason for this phenomenon is a fundamental imbalance in their training (Pi et al., 2024; Sun et al., 2024; Yu et al., 2023, 2024). In this paper, we evaluate a suite of contemporary MLLMs on the Vaani and Places audio caption datasets to assess the alignment between model- and human-generated captions.

3 Datasets and Background

THE VAANI DATASET: The Vaani dataset (VAANI, 2025) is a large-scale, open-source, multilingual, multimodal corpus comprising approximately

27,751 hours of spontaneous, image-prompted speech collected from 143K speakers across 145 Indian districts. It includes descriptions of 258K images spanning 103 languages, of which 1,187 hours are manually annotated. This work focuses on the Hindi subset, the largest monolingual component of the transcribed data, containing 594 hours of speech. We use its test split, consisting of 9,888 unique image–transcript pairs from 3,288 speakers.

For each image, participants provide spoken descriptions in their native language or dialect. Each audio recording is accompanied by metadata specifying the language, demographic attributes (gender, state, and district), and a verbatim transcription. The manually annotated transcripts serve as the human captions in our bidirectional evaluation framework (Section 4.2). VAANI captures natural, spontaneous visual descriptions, including hesitations or incomplete phrasing. This makes it particularly well suited for evaluating image captioning models under realistic conditions, where system outputs can be directly compared against speech-derived human captions. An example is shown in Table 1.

PLACES AUDIO CAPTIONS (HINDI): We further evaluate our framework on the Hindi subset of the Places Audio Captions dataset (Harwath et al., 2018), built on the Places205 scene-centric image corpus (Harwath et al., 2016). The dataset comprises free-form, unprompted spoken descriptions collected by asking participants to describe objects and scenes in an image, covering 85,480 images. These natural spoken captions are automatically transcribed using the Google Automatic Speech Recognition (ASR) system and are treated as human captions in our bidirectional evaluation. Examples for the dataset could be found in Table 4 (Appendix Section A.4) .

Table 1: Reference images and human-generated captions from Vaani dataset

Reference image	Transcript
 Specific_01	<noise> Hindi: इस फोटो {photo} Hindi: में सामने की तरफ एक लड़का दिखाई दे रहा है जो स्कूटर {scooter} Hindi: पे पर बैठा हुआ दिखाई दे रहा है जिसने ब्लू {blue} Hindi: कलर {colour} Hindi: की टी-शर्ट {t-shirt} </noise>
 Generic_03	<noise> Hindi: यहां पाये मशीन {machine} Hindi: लगी हुई हैं अंदर और आदमी कम कर रहे हैं और मशीनों का रंग सफेद हैं और आदमी कल हेला हैं।</noise>

4 Methodology

4.1 DCScore metric

Ye et al. (2025) proposed the DCScore, a novel metric designed to evaluate hallucinations and factual correctness. This involves the following steps:

1. Decomposition: Generated captions and ground-truth captions are decomposed into the standalone facts or self-sufficient units, referred to as Primitive Information Units (PIUs) (Ye et al., 2025), using MLLMs. Human transcripts are broken down into a set of PIUs, denoted as $T = \{t_1, t_2, \ldots, t_M\}$, where M is the number of extracted units. Similarly, model-generated captions are decomposed into a set $P = \{p_1, p_2, \ldots, p_N\}$, where N represents the number of extracted units.

2. Matching: An LLM is prompted to determine whether each primitive unit in $p_i \in P$, from the generated captions, is either explicitly stated or logically inferable from a corresponding unit $t_j \in T$ in the human transcript. The alignment between generated and human captions is defined as $Q = P \cap T$, where Q represents the set of overlapping PIUs.

3. Verification: The PIU in the generated caption is verified against the input image using an LLM. The verification process evaluates the accuracy of each unit p_i in the generated captions P by directly referencing the corresponding image. Following the original DCScore methodology, we employ GPT-4.1 (OpenAI, 2025), with stable API access, to guarantee that our evaluation metric is deterministic and fully reproducible across studies.

Evaluation Metrics

From the model-generated set P, the human transcript set T, and their overlap Q, precision, recall and F1 scores are defined to evaluate caption quality.

Precision score s_p measures the proportion of

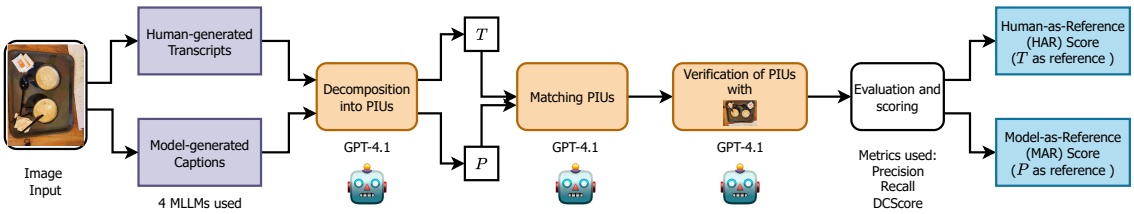


Figure 1: Proposed Bidirectional framework to assess the alignment between model- and human-generated captions.

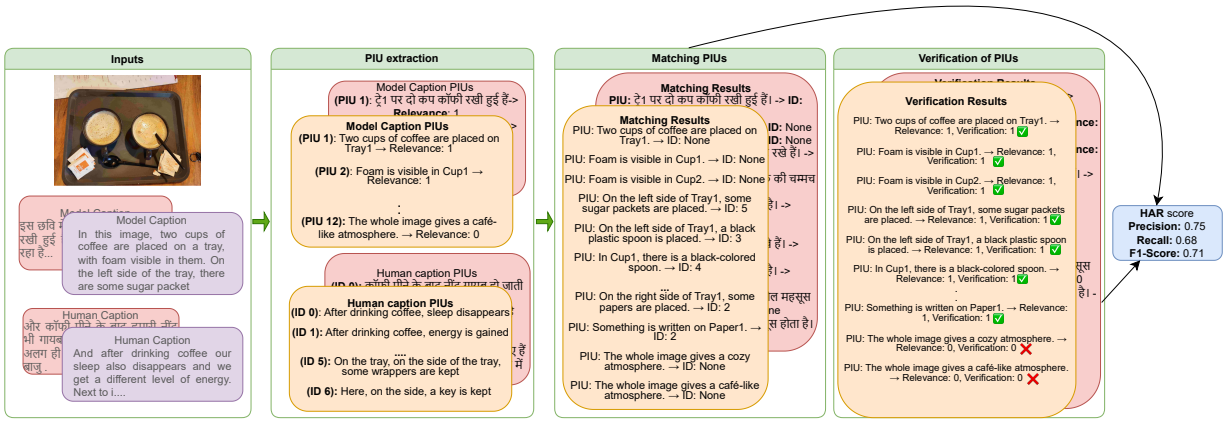


Figure 2: Illustration of Human-as-Reference (HAR) evaluation for a single data point. The corresponding MAR evaluation is depicted in Figure 6.

correct PIUs among those produced by the MLLM:

$$s_p = \frac{|P_{true}|}{|P|} \tag{1}$$

where $P_{true} = \{p_i \mid p_i \in P,\ p_i \text{ is correct}\}$ denotes the subset of correct units in P.

Recall score s_r measures the proportion of human transcript PIUs that are either directly aligned or correctly produced by the MLLM:

$$s_r = \frac{|Q| + |P_{true} \setminus Q|}{|T| + |P_{true} \setminus Q|}, \tag{2}$$

The overall quality of captioning is measured using the **F1 score** s_f, which is the harmonic mean of precision (s_p) and recall (s_r).

4.2 Proposed Bidirectional Evaluation Framework

Inspired by the DCScore (Ye et al., 2025), we propose a bidirectional framework(Figure 1) containing the following parts.

1. **Human-as-Reference (HAR) Score:** Here, the model captions form the target text, which are evaluated against the human caption reference.
2. **Model-as-Reference (MAR) Score:** Here, the human-captions form the target text, which are evaluated against the model reference.

Figures 2 and 6 (Appendix A.1) provide a concrete example of our bidirectional scoring framework for a single image. For clarity, the results in this figure are translated into English; the original Hindi text is shown as a red underlay.

In the **Human-as-Reference (HAR) evaluation** (Figure 2), the subjective PIUs from the model, such as Hindi: "पूरी छवि में कैफे जैसा माहौल महसूस होता है"[1], are assigned a relevance score of 0, as they are not direct visual facts. The framework then quantifies the alignment to produce the precision (0.75), recall (0.68), and F1 (0.71). Conversely, the **Model-as-Reference (MAR) evaluation** (Figure 6) highlights how human descriptions can include inferential statements that are not visually grounded in the image. The example highlights the need for bidirectional evaluation. Thus, we move beyond the DC score and adopt the HAR/MAR framework to allow for a more flexible evaluation setting, particularly

[1] *The entire image gives the feeling of a café-like atmosphere.*

in scenarios where reference captions do not constitute a single canonical or gold-standard description of the visual content, as in the Vaani dataset.We further compare HAR/MAR with standard reference-based metrics by analyzing their correlation with human judgments (Appendix **??**). More details of the bidirectional evaluation is provided in Appendix A.2.

4.3 Experimental Setup

Models Evaluated: Our experimental setup evaluates four MLLMs, selected across sizes and various model families (Table 2). Each model was tasked with generating a caption for every image in our dataset using the standardized prompt: *"Describe the image in comprehensive detail as a single paragraph in Hindi."*

Table 2: Overview of the MLLMs used in our bidirectional evaluation framework.

Stage	MLLM	Release
Caption Generation	GEMINI 2.5 PRO[2] [No Thinking]	2025
	GPT-4O[3] [No thinking]	2024
	GEMMA-3-12B[4]	2024
	LLAMA-4-SCOUT-17B-16E-INSTRUCT[5]	2025
Decomposition, Matching & Verification	GPT-4.1[6] [No thinking]	2025

5 Results

We evaluate results along four aspects:

1. **Bidirectional scoring**: We examine both Human-as-Reference (HAR) and Model-as-Reference (MAR) evaluation scores to capture asymmetries of evaluation.

2. **Statistical validation:** We assess the robustness of performance differences using unpaired t-tests (Welch, 1947), comparing each model against the highest-F1 baseline for significance.

3. **Error Quantification:** We define:

   (a) Hallucination Rate as $1 - \text{Precision}$, i.e., the proportion of generated PIUs that are not verified against the reference caption and Image.

   (b) The Omission Rate: the proportion of reference PIUs that are not captured by the reference caption.

$$\text{Omission Rate} = \frac{N_{\text{reference}} - N_{\text{matched}}}{N_{\text{reference}}}$$

4. **Sample difficulty:** To further examine the differences at category-level, we analyze sample difficulty by bucketing each image into Easy, Medium, or Hard based on a global HAR/-MAR performance threshold.

Results for the Vaani dataset are reported in the main text, while the corresponding plots and tables for the Places audio caption dataset are provided in Figures 13, 14 (Appendix A.5.1).

5.1 Model-wise Analysis

Figure 3 for the Vaani dataset (Figure 13 (Appendix A.5.1) for the Places Audio Captions dataset) present model-wise performance under the bidirectional scoring framework for a random set of 200 images. Results indicate that GEMINI-2.5-PRO achieves the highest mean F1 score in the HAR setting but comparatively lower F1 in MAR. A high MAR score reflects more comprehensive captions, with stronger coverage (Recall), as further illustrated in Figure 12 (Appendix A.5.1) for the Vaani dataset and Figure 14 (Appendix A.5.1) for Places Audio Captions. In contrast, GPT-4o shows the opposite trend, with a comparatively lower F1 score in HAR, but a high F1 score in MAR, suggesting that its captions are less exhaustive yet more closely aligned with human-generated captions compared to the other models. To contextualise these results, precision-recall trade-offs are further analyzed in Figure 12 (Appendix A.5.1) [Places Audio Captions-Figure 14 (Appendix A.5.1)].

For statistical validation across both datasets, Vaani and Places Audio Captions, each model was compared with GEMINI-2.5-PRO in the HAR evaluation pipeline and GPT-4o in the MAR evaluation pipeline, using unpaired t-tests. In the HAR direction, the differences in F1-score between GEMINI-2.5-PRO and the other models (GEMMA-3-12B, GPT-4o, and LLAMA-4-SCOUT) were statistically significant ($p < 0.05$), demonstrating that GEMINI-2.5-PRO is the best performing model statistically. In

[2] https://deepmind.google/models/gemini/pro/
[3] https://openai.com/index/hello-gpt-4o/
[4] https://huggingface.co/google/gemma-3-12b-it
[5] https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E
[6] https://openai.com/index/gpt-4-1/

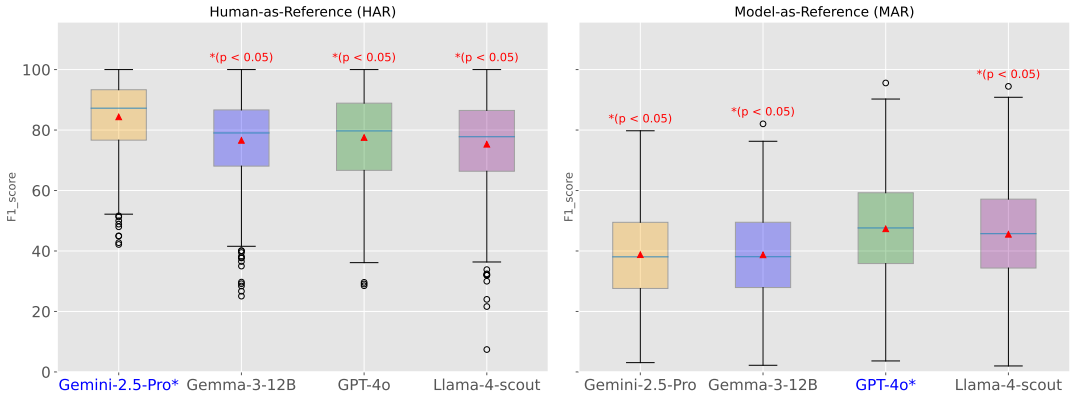


Figure 3: Boxplot visualization of the DCScore (HAR and MAR) of 200 random image samples, across all models. The reference (highest F1 score) model is highlighted in blue on x-axis and models with $p < 0.05$ are significantly different from the reference.

the MAR direction, GPT-4o significantly outperformed the alternative models.We further validate the robustness of these model-wise trends by repeating the HAR and MAR evaluation on an English translation of the Vaani dataset, observing consistent model rankings across HAR and MAR settings (Appendix Section A.5.3).

Table 3 reports hallucination and omission rates across the Vaani and Places Audio Captions datasets, providing insights into model error profiles beyond F1 differences. In the HAR direction, hallucination rates remain consistently low (2.4%–12.4%) across models. GEMINI-2.5-PRO (lowest hallucination and omission) and GPT-4o exhibit the strongest control, with hallucination rates of approximately 2.4%–3.7% on Vaani and 3.7%–4.2% on Places, while GEMMA-3-12B shows elevated hallucination (12.4% on Vaani; 8.3% on Places), indicating a tendency to exaggerate details. Despite low hallucination, omission rates are substantial (28.4%–50.6%), with GPT-4o exhibiting higher omission (50.6% on Vaani; 30.7% on Places), while GEMINI (43.5% on Vaani; 28.4% on Places) and GEMMA (48.1% on Vaani; 30.4% on Places) achieve comparatively better coverage.

In the MAR direction, omission rates are high (60.4%–85.1%), reflecting the limited coverage of human-generated captions relative to more exhaustive model descriptions across both datasets. GEMMA is the most omission-prone (71.1% on Vaani; 82.4% on Places), while GPT-4o exhibits the best balance between omission and hallucination. These elevated MAR omission rates indicate that human annotations typically capture only 20.0%–30.0% of the PIUs in an image. In contrast, models omit substantially fewer PIUs, with omission rates around 40.0% on Vaani and 30.0%–35.0% on Places, indicating broader visual coverage. Notably, this effect is amplified on the Places dataset, where MAR human omission rates are consistently high across all models.

Table 3: Hallucination and omission rates (in %) on the Vaani dataset HAR and MAR settings. Corresponding values from the Places Audio Captions dataset are shown in red within parentheses.

Set.	Model	Hall. (%)	Omis. (%)
HAR	Gemini-2.5-Pro	2.4 (3.7)	43.5 (28.4)
HAR	Gemma-3-12B	12.4 (8.3)	48.1 (30.4)
HAR	GPT-4o	3.7 (4.2)	50.6 (30.7)
HAR	Llama-4-scout	6.8 (3.8)	49.8 (36.6)
MAR	Gemini-2.5-Pro	24.6 (28.1)	67.4 (85.1)
MAR	Gemma-3-12B	24.4 (30.5)	71.1 (82.4)
MAR	GPT-4o	26.4 (28.2)	60.4 (80.4)
MAR	Llama-4-scout	24.1 (29.4)	63.3 (81.4)

From these observations, we conclude that there is a consistent divergence between human- and model-generated captions. Humans demonstrate strong visual selectivity, i.e., they describe only a subset of the image, focusing on salient regions while omitting secondary details. Models, by comparison, generate captions that are denser and more exhaustive, often attempting to cover the entire image. This discrepancy reflects the divergent objectives - humans prioritise communicative sufficiency, whereas models are optimized for coverage.

5.2 Category-wise Analysis

Category Selection: The unique set of 9,888 images from the Vaani Hindi test set was further used for categorization. The prompt and strategy used for this classification is provided in Appendix

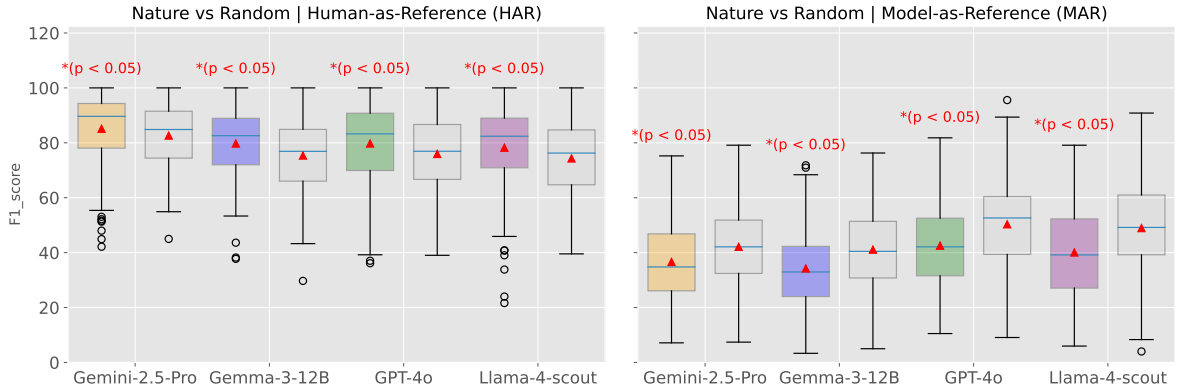


Figure 4: Boxplot visualization of the DCScore (HAR and MAR) across all models for nature and random categories. For nature, all models are significantly different ($p < 0.05$) from the random category. Grey colour represents the random category across all models, while the nature category is shown in a distinct colour for each model.

A.3. Examples are provided in Appendix Section A.4. The categories are:

1. **Commercial & Retail:** Scenes of markets, vendors, and trade.
2. **Religion & Culture:** Religious ceremonies, festivals, or cultural heritage.
3. **Nature & Landscape:** Natural elements and wildlife.
4. **Education & Learning:** Academic or instructional settings.
5. **Infrastructure & Transport:** Buildings, public works and modes of transit.

Figure 4 reports the statistical significance results for the nature category relative to the random reference. In the HAR setting, across all four models, the F1 scores for nature are significantly higher than random images ($p < 0.05$), indicating that models consistently describe natural scenes more effectively than random ones. In the MAR setting, F1 scores are significantly higher than the random baseline ($p < 0.05$). This asymmetry reflects the fact that humans are more selective in certain categories. Figure 16 in Appendix A.5.1 shows the F1 score across five categories for all four models.

Hallucination and Omission Rates: Referring to the HAR/MAR hallucination and omission rates in Section 5.1, HAR omission rates broadly follow visual complexity. Commercial and Infrastructure categories exhibit the highest omissions (≈52% and ≈49%), followed by Culture (≈48%), while Random and Educational categories are slightly lower (≈48% and ≈44%). In contrast, MAR omission rates are higher overall (≈60%–≈75%), with Educational and Nature scenes showing the highest omissions (≈72%). This reversal in MAR omissions indicates that models tend to produce more verbose descriptions than humans in these categories.

5.2.1 Thematic Category Difficulty Distribution

To further examine category-level differences, we analyze the difficulty of individual samples by bucketing each image into *Easy*, *Medium*, or *Hard* based on a global HAR/MAR F1 threshold (Figure 18, Appendix A.5.1). This analysis reveals a clear pattern: the Nature and Educational categories contain the highest proportion of *Easy* samples and the lowest proportion of *Hard* samples in HAR.

As illustrated by a Nature-category example in Figure 17 (Appendix A.5.1), the model caption captures high-level scenic elements (e.g., reservoir, vegetation, reflections, sky), which align well with transcript PIUs such as “pond” and “plants”, yielding a strong HAR score (F1 = 0.98). In contrast, in the MAR direction, transcript PIUs such as “water flowing abundantly” are absent from the model caption and are not directly inferable from the image, leading to failed verification and a substantially lower MAR score (F1 = 0.25). While elevated HAR omission is a common challenge in visually dense scenes, the accompanying increase in HAR hallucination is model-dependent. This effect is most pronounced for GEMMA, which exhibits comparatively high hallucination and omission rates in the Commercial category. A representative example is shown in Figure 22 (Appendix A.5.1).

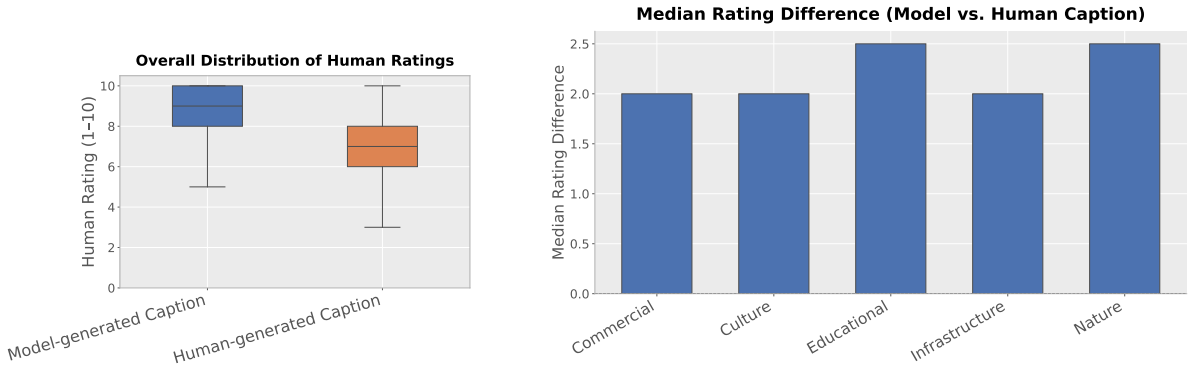


Figure 5: Figure (a) box-plot visualization of comparative human rating for model and human-generated captions and (b) median difference in human ratings between model-generated (Gemini-2.5-Pro) and human-generated captions across five semantic categories.

5.3 Human Judgment of Caption Quality

To further understand human preferences, we conducted a human evaluation study focused on caption coverage and comprehensiveness. Participants were asked to rate both human- and model-generated captions on a 1–10 scale based on how well they covered visible details while excluding irrelevant information, ignoring linguistic fluency. To ensure a fair comparison, model-generated captions were constrained to match the median length of the human-generated captions. We randomly selected 22 images, each paired with a human caption and a GEMINI-2.5-PRO caption, with the presentation order randomized across images. In total, 16 native Hindi speakers independently rated all captions.

Key takeaways for results plotted in Figure 5 :

1. **Model captions are more preferred:** Figure 5(a) shows the rating distribution of model and human-generated captions. This reflects the greater descriptive consistency of GEMINI-2.5-PRO, indicating that the captions capture comprehensive details while presenting them in a more fluent manner. The reduced preference to human captions also mirrors the DCScore analysis, where higher F1 values of HAR over the MAR scores were observed.(Figure 3).

2. **Certain categories accentuate the human-model captioning differences:** In Figure 5(b), we delve into the preferential pattern seen in Figure 5(a), by plotting the median difference between the ratings given by the human subjects to the model-generated captions and human-generated captions across the 5 semantic categories. The performance difference is the most pronounced for the 'nature' and 'educational' categories, a result that corroborates our category analysis using DCScore (Sec. 5.2.1). The human-ratings reinforce the choice of the DCScore as the tool for the analysis reported in this work.

Although evaluators were instructed to assess content rather than linguistic fluency, several noted that human captions were abrupt or contained colloquial variations. In contrast, model-generated captions exhibited a more polished and structured style, which likely contributed to their higher consistency scores.

6 Conclusion

We present a bidirectional evaluation framework that jointly assesses human- and model-generated captions across 4 SOTA MLLMs. Our analysis reveals a fundamental asymmetry in captioning behavior: humans exhibit selective attention, describing salient aspects of an image while omitting secondary details, whereas models tend toward more exhaustive coverage. Among the evaluated models, GEMINI-2.5-PRO produces the most comprehensive captions with minimal hallucination and omissions, while GPT-4o exhibits the most human-aligned behavior, generating more selective and specific descriptions. Such model-wise performance trends remain dataset and language agnostic. Together with human judgment results, our findings highlight an important design trade-off between coverage-oriented safety and human-aligned selectivity, and suggest that future models should explicitly account for this balance when targeting human-centric applications.

7 Limitations

Our analysis primarily focused on understanding the alignment between natural, informal human captions and model-generated captions in the Vaani dataset, quantified through omission and hallucination rates within a bidirectional evaluation framework.

In the Vaani dataset, It is important to note that models have the advantage of generating captions with a length comparable to those produced collectively by multiple humans, whereas each human caption is influenced by the need to caption multiple images in a sequential manner. Although there was no explicit push to speed up the human captioning process, the implicit behavior to caption as many images, may have influenced the human behavior of omitting details of the image. Additionally, our analysis is limited to two datasets, Vaani and Places Audio Captions, and primarily focuses on a single language (Hindi). While both datasets capture spontaneous, speech-based image descriptions, they differ in collection protocols, transcription quality, and cultural context, which may influence the observed selectivity patterns.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023a. Qwen technical report. *Preprint*, arXiv:2309.16609.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023b. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. 2023. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

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

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and 1 others. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Elaine Giles, Karalyn Patterson, and John R Hodges. 1996. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*, 10(4):395–408.

Zenzi M Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological science*, 11(4):274–279.

David Harwath, Galen Chuang, and James Glass. 2018. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. *Preprint*, arXiv:1804.03052.

David Harwath, Antonio Torralba, and James Glass. 2016. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Sen He, Hamed R Tavakoli, Ali Borji, and Nicolas Pugeault. 2019. Human attention in image captioning: Dataset and analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8529–8538.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Bras, and Choi Yejin. 2021a. Clipscore: A reference-free evaluation metric for image captioning. pages 7514–7528.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021b. CLIPScore: a reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528.

Falk Huettig, Joost Rommers, and Antje S Meyer. 2011. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2):151–171.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Hyunjong Kim, Sangyeop Kim, Jongheon Jeong, Yeongjae Cho, and Sungzoon Cho. 2025. Expert: An explainable image captioning evaluation metric with structured explanations. *arXiv preprint arXiv:2506.24016*.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. *arXiv preprint arXiv:2406.06004*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

OpenAI. 2024. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`.

OpenAI. 2025. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. *ArXiv*, abs/2403.08730.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6914–6924.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.

Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2024. Describing images fast and slow: Quantifying and predicting the variation in human signals during visuo-linguistic processes. *arXiv preprint arXiv:2402.01352*. Studies onset times, durations, variation in human behavior signals.

VAANI. 2025. Vaani: Capturing the language landscape for an inclusive digital india (phase 1). https://vaani.iisc.ac.in/.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

B. L. Welch. 1947. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *EMNLP*, pages 444–454.

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. 2025. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *The Thirteenth International Conference on Learning Representations*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2023. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. 2024. Meacap: Memory-augmented zero-shot image captioning. pages 14100–14110.

A Appendix

A.1 Bidirectional DCScore Evaluation pipeline

Following caption generation, these captions and the corresponding human transcripts were processed through a three-stage evaluation pipeline orchestrated by GPT-4.1 [No thinking] as can be seen in Figure 1. First, in the Decomposition stage (Section 1), Both Model caption and reference caption are broken into Primitive Information Units (PIUs) as per the prompt in Appendix A.2. Next, the Matching stage (Section 2) uses the prompt in Appendix A.2 to map each model PIU to a reference caption PIU ID or "None" if no correspondence is found. In parallel, the Verification stage (Section 3) assesses the factual correctness of each model PIU against the source image or the reference caption, yielding a binary score of "1" for correct and "0" for incorrect (see Appendix A.2). The entire HAR-MAR evaluation process with the intermediate results is illustrated in Figure 2(HAR)(main text) and Figure 6(MAR).

For the Model-as-Reference (MAR) evaluation, we maintain methodological consistency by using the same set of prompts as in the HAR pipeline. This is achieved by simply reversing the inputs: the human transcript is treated as the 'predicted caption' to be evaluated, while the corresponding model-generated caption serves as the 'reference caption' ground truth. As illustrated in our example figures (Figure 2(Main text) and 6), this role reversal is seamless, as the set of PIUs extracted from each text remains identical regardless of its role in the evaluation pipeline.

To fairly handle many-to-one mappings where multiple predicted PIUs may correspond to a single Transcript PIU-DCScore employs a weighting scheme. The contribution of each correct, predicted PIU to the recall score is weighted by $1/N$, where N is the count of prediction PIUs mapped to a single Human-generated caption PIU.

Captions generated from human descriptions of images exhibit substantial variability in both content and level of detail. Some captions are highly elaborate incorporating contextual information, others are minimal, strictly grounded in the visible scene. Such variability introduces inconsistencies, with descriptions that may be underspecified or include information not directly supported by the image. For example, a human caption might describe construction workers as"working very hard" or "They have been working for a few days, and the work has just begun," as shown in Figure 11. These are subjective or temporal assessments rather than literal observations grounded in the image. Consequently, while this reduces precision in the MAR direction – since not all human-generated

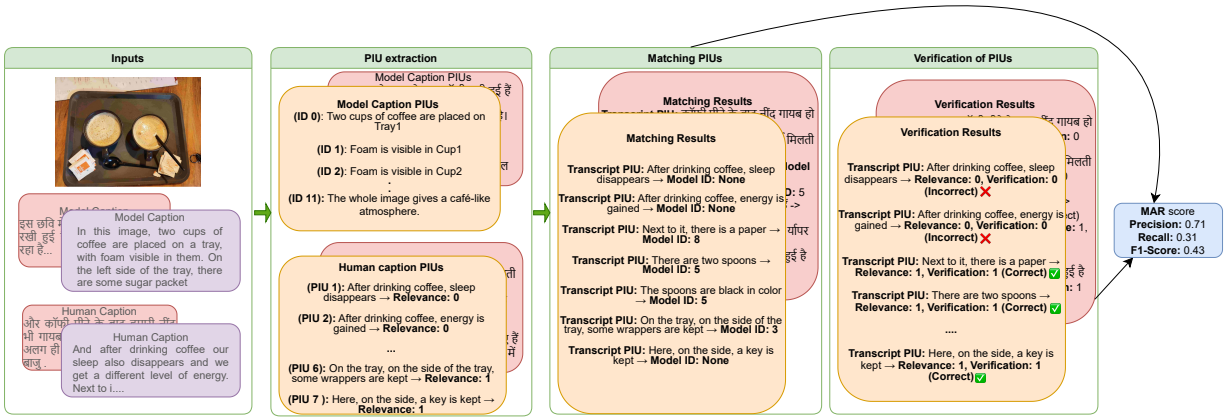


Figure 6: Illustration of Model-as-Reference (MAR) evaluation for a single data point

PIUs can be verified against model captions/image – we still observe a relatively higher MAR Recall, hence a higher F1 score in this case. These deviations show why human captions cannot always be treated as flawless gold references. Since human-generated captions are not completely relaible, we generate machine captions using MLLMs for images in the Vaani dataset and propose a bidirectional human–model evaluation framework that compares human- and model-generated captions to assess the credibility of human annotations.

A.2 Prompts used in the DCScore framework

Prompt 1 for Decomposition into PIUs

You are a linguistic expert in extracting primitive information units in the given image caption which is in HINDI and related to the image. Consider the image content when extracting PIUs. In specific, "primitive information units" refer to the smallest standalone pieces of information that collectively represent the entire meaning of the sentence without losing any detail, which typically describe various properties of the visual elements in an image. The primitive information unit should only contain ONE primary element. When extracting primitive information units from image caption, it is useful to assign unique identifiers to the primary objects or entities being discussed. This will help in maintaining clarity and preventing confusion, especially when there are multiple similar objects or entities. For example, if the caption mentions two cats, you can assign unique identifiers such as "cat1" and "cat2" to distinguish them. Besides, for each attribute, you should also assign the identifier to the object it belongs to. Meanwhile, for spatial relationships, you can assign the identifier to the object that is the subject of the relationship in the primitive information unit. For each primitive information unit, you should also need to justify whether the primitive information unit directly describe the image or not. IMPORTANT: Please extract ALL of the primitive information units in the image caption. DO NOT omit any information! Please make sure the identifiers are uniform and only in hindi , with maybe the attached 1,2,3.. like Hindi: मेज1, Hindi: फूल1, Hindi: चटनी1. The output should be a list of dict ["fact": [PRIMITIVE INFORMATION UNIT], "identifier": [UNIQUE ID], "relevance": 1/0, ...] The output of the PIUs should only be in HINDI strictly as input. dont need any justification for the answers just in the explained format.
»> Caption:

Prompt 2 for matching

You are now a visual-linguistic expert in matching two set of primitive information units generated from A caption which is VLM generated and another Transcript(Human). You will be received a set of predicted VLM primitive information units across a variety of categories and a set of Transcript primitive information units (ground truth).

The set of primitive information units is represented as a list of dict [”fact”: [PRIMITIVE INFORMATION UNIT], ”identifier”: [UNIQUE ID], ...] within JSON format. In addition, each primitive information unit in the Transcript set would be accompanied with a unique ”id” to identify the Transcript primitive information unit. To match primitive information units (PIUs) from a predicted VLM set with a Transcript set:

1. Preliminary Review: Conduct an initial review of both sets of primitive information units, considering all primitive information units. Understand the details and context presented within each primitive information unit.

2. Inferring Identifier Mappings: Closely examine both sets to deduce potential correlations and mappings based on the content of the primitive information units. Determine if there are any unique identifiers or descriptors that hint at matching entities between the sets. For example, ”cat0” in the predicted VLM set’s primitive information units may be mapped to ”cat1” in the Transcript set’s primitive information units. Consider the attribute and spatial relation in both sets for possible mapping.

Please note that there might be some attribute and spatial errors when mapping the objects. Try find the most similar mapping if exists (not need exact matching). If no Transcript primitive information unit matches, simply set matched Transcript id to ”None”. **IMPORTANT**: Please consider each primitive information unit in the set individually, and MUST NOT omit any primitive information units from the predicted VLM set. You should only output the matching results which will be formatted as a list of dict as [“fact”: [PRIMITIVE INFORMATION UNIT],
“identifier”: [UNIQUE ID],
“matched_transcript_id”: [CORRESPONDING Transcript ID], ...] in JSON format. The “identifier” would be optional, if the item in the fact has already been identified with ids as illustrated in the predicted VLM primitive information units. For key named “matched_transcript_id”, the value of “matched_transcript_id” should be the corresponding “index“ of the primitive information unit in the Transcript set. For the primitive information unit in the predicted VLM set which cannot be matched with any Transcript primitive information unit, set the value of “matched_transcript_id“ to “None”. You must produce an output for each VLM predicted primitive information unit, attempting to match it against the transcript set.
»> Set of VLM predicted Primitive information units:
»> Transcript Set of Primitive information units:
»> Matching Result:

Prompt 3 for verification

You are an extraordinary visual-linguistic expert in verifying the correctness of a set of primitive information units given the image and the corresponding reference caption. The set of primitive information units are extracted from a paragraph of machine-generated image caption of that image. The set of primitive information units is represented as a list of dict [”fact”: [PRIMITIVE INFORMATION UNIT], ”identifier”: [UNIQUE ID], ...] within JSON format. The identifier is unique and to identify the primary objects or entities being discussed. This will help in maintaining clarity and preventing confusion, especially when there are multiple similar objects or entities. For example, if the caption mentions two cats, we would assign unique identifiers such as ”cat1” and ”cat2” to distinguish them. Besides, for each attribute, it also assigned the identifier to the object it belongs to. Meanwhile, for spatial relationships, it assigned the identifier to the object that is the subject of the relationship in the primitive information unit. You should first go through all of the primitive information units, and understand the details

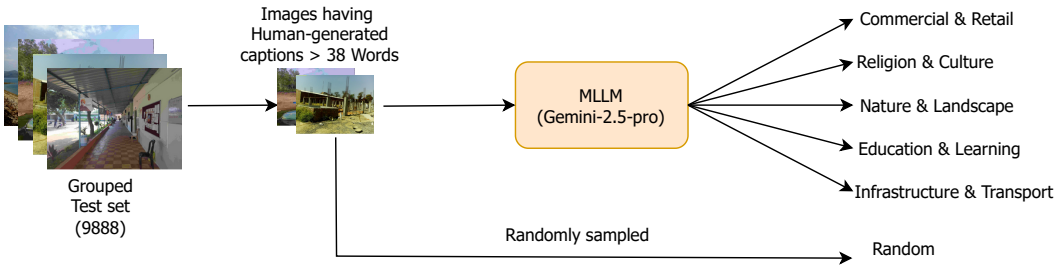


Figure 7: The Pipeline illustrating the selection of images across different categories. we had to first do a multi-fold passes of 1000s of images, This process revealed the five primary categories detailed. We selected only those that were unambiguously classified into a single semantic category

and context presented within each primitive information unit. Then you need to verify the correctness of each individual primitive information units by asking yourself: Statement: "[PRIMITIVE INFORMATION UNIT]" Does the statement correct according to image or reference caption? The output for the predicted VLM set of primitive information units should be formatted as a list of dict as ["fact": [PRIMITIVE INFORMATION UNIT], "identifier": [UNIQUE ID], "verification": 1/0, ...] in JSON format, where 1 represents the fact is correct and 0 represents the fact is incorrect. Other keys in the dictionary are the same as the input. The "identifier" would be optional, if the item in the fact has already been identified with ids as illustrated in the input. Output only the json in the given format, no hallucinations or explanations or justification.
»> Reference Caption:
»> Primitive Information Units:

A.3 Category classification

Prompt used for category classification through Gemini-2.5-Pro

You are an expert image classifier with a deep understanding of Indian cultural contexts. Your task is to analyze an image and classify it based on its relevance to the categories defined below. For each image, identify the most appropriate category. In most cases, there should be a single dominant category that captures the main theme of the image. Only in rare cases—when two categories are both clearly represented—may you assign two categories. Even then, only one can be rated as High (the central theme), while the other must be rated as Medium or Low. For every category assigned, provide a rating (High, Medium, or Low) based on its prominence in the image, along with a brief justification. —
Categories & Definitions
1. Commercial & Retail: Scenes centered on buying, selling, or trade. Examples: Markets, street vendors, kirana stores, malls, commercial signage. 2. Religion & Culture: images showcasing religious practices, cultural heritage, or festivals. Examples: Temples, mosques, church, pujas, traditional attire (saree, kurta), cultural performances. 3. Nature & Landscape: Outdoor scenes dominated by natural elements. Examples: Mountains, rivers, beaches, forests, gardens, wildlife. 4. Education & Learning: Settings related to academic or instructional activities. Examples: Schools, classrooms, students in uniform, libraries, teachers. 5. Infrastructure & Transport: Man-made public works and modes of transit. Examples: Roads, bridges, railway stations, airports, buses, trains, auto-rickshaws.
—
Rating Guidelines When assigning a rating (High, Medium, Low) to a category, carefully consider the intensity and dominance of that theme in the image: 1. High → The category is the clear central theme of the image, visually dominant and unambiguous. 2. Medium → The category is present and noticeable, but not the primary focus. 3. Low → The category is only weakly or marginally represented, secondary to other themes. Do not default to "High" whenever a category is present. Reserve High only for strong, central cases. Use Medium or Low where the category is visible but less dominant.
—
Output Format Respond with a JSON array of objects. Each object represents a classified category and must include category, rating, and reasoning. The array will contain one object, or in cases of significant overlap, two objects.
- If the image does not contain significant elements from any category, classify it as "Normal" with a rating of "N/A". - Do not make assumptions or hallucinate; categorize based only on visual evidence.
Example Output (Single Category):

```
[
  {
    "category": "Infrastructure &
        Transport",
    "rating": "High",
    "reasoning": "The image is a wide
        shot of a bustling railway
```

```
        station,
    with trains and passengers as the
        central focus."
  }
]
```

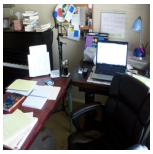
Example Output (Multiple Categories):

```
[
  {
    "category": "Religion & Culture",
    "rating": "High",
    "reasoning": "The central subject
        is a group of people in vibrant,
    traditional attire participating in
        a religious procession."
  },
  {
    "category": "Commercial & Retail",
    "rating": "Medium",
    "reasoning": "The procession is
        taking place in a crowded market
    street, with numerous shops and
        street vendors clearly visible
    in the background."
  }
]
```

A.4 Places Audio captions Examples and Vaani Category Examples

The examples for Places Audio Captions Dataset can be seen in Table 4 and the category wise samples used for Categrory-wise analysis of the Vaani dataset can be found in Figures 8 and 9.

Table 4: Example images and ASR-transcribed Hindi captions from the Places Audio Captions dataset.

Image	asr_text (Hindi)
	एक गाड़ी के अंदर कुछ लोग काम कर रहे हैं और इन सबने एक जैसी वर्दी पहनी हुई है
	यह काफी सामान लेकर यहां खड़ा नजर आ रहा है और नीले दरवाजे के आगे खड़ा है वह बाहर
	क्या आप देख सकते हैं मैच में कागज है लैपटॉप दिखाई दे रहा है कुर्सी है

A.5 Additional Results

A.5.1 Word-Length-Constrained Model Performance Analysis

In parallel to the word-length constraints put on model captions for human judgment, we also examine how trends and variations occur when model captions are constrained to the median length of human-generated captions.We would also like to point out further that the aggregation of multiple speaker captions for a single image does not yield the strong overlap one might intuitively expect—in fact, the overlap is considerably limited. For example, if 5 people described an image but only one mentioned a given PIU, its consensus is 1. If two people mentioned it, the consensus is 2, and so on. These PIU-level consensus counts were averaged at the image level and subsequently aggregated by category. The distributions shown in the figure 10b indicate that consensus across speakers is generally low, hovering near 1.25, highlighting that even when multiple individuals describe the same image, In most cases they tend to focus on different subjects. Notably, Nature images exhibit somewhat higher variability in comparison to other categories which reflect limited overlap. This effectively does away with the argument of aggregation of the same subjects over many speakers.

Looking at model-wise comparisons on Random images under these constraints, we see the same overall trends as in the unconstrained setting, though the differences across models are less pronounced, likely owing to smaller variations in caption length. The performance difference in both HAR and MAR direction for Gemini-2.5-Pro could be found in figure 10a. In the HAR direction, Gemini-2.5 Pro continues to lead, followed by GPT-4o and Gemma-3-12B. In the MAR direction, GPT-4o maintains an edge, while Gemini and Gemma cluster lower. Notably, LLaMA has been excluded from the constrained caption plots due to its gross under performance, with a majority of outputs exhibiting hallucinations as can be seen in the Hallucination and Omission rates plots in Figure 20. Gemma-3-12B continues to exhibit the highest hallucination rate among the competitive models. The Human omission rates (MAR) also remain significantly higher than the Model omission rates (HAR), solidifying the claim that humans demonstrate strong visual selectivity. Among the models, LLama-4-17B displays the highest hallucination rate in the constrained captioning setting,

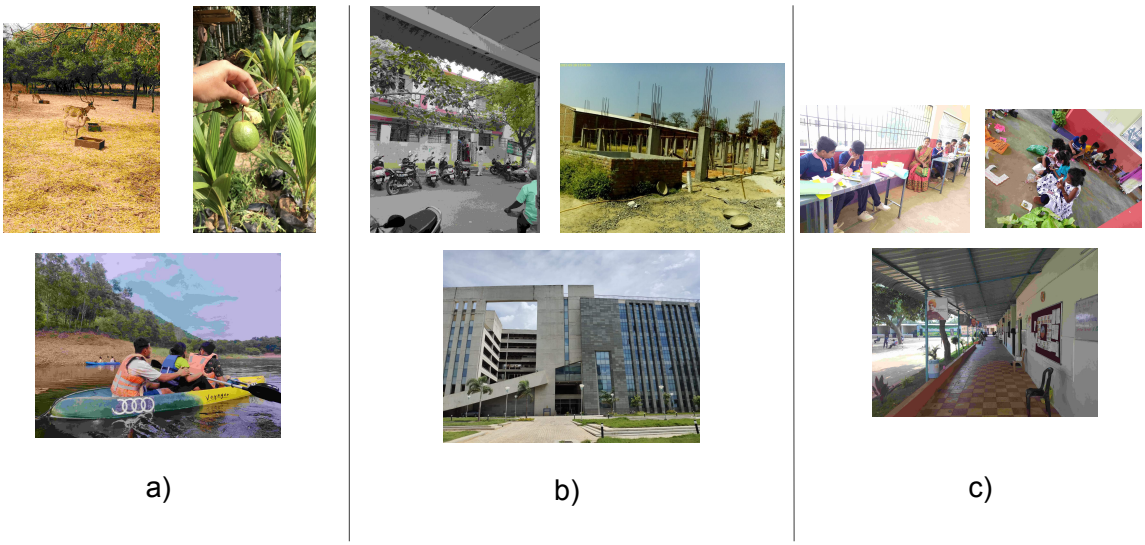


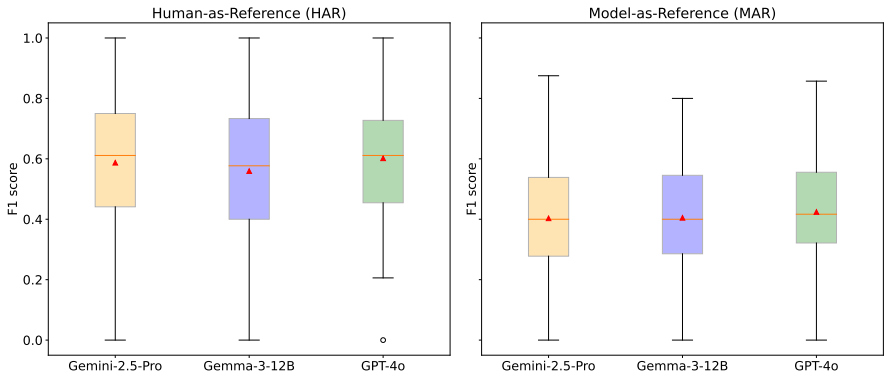
Figure 8: image examples from a) Nature & Landscape, b) Infrastructure & Transport and c) Education & Learning categories during sub-categorization.

Figure 9: image examples from d) Commercial & Retail, e) Religion & Culture categories during sub-categorisation.

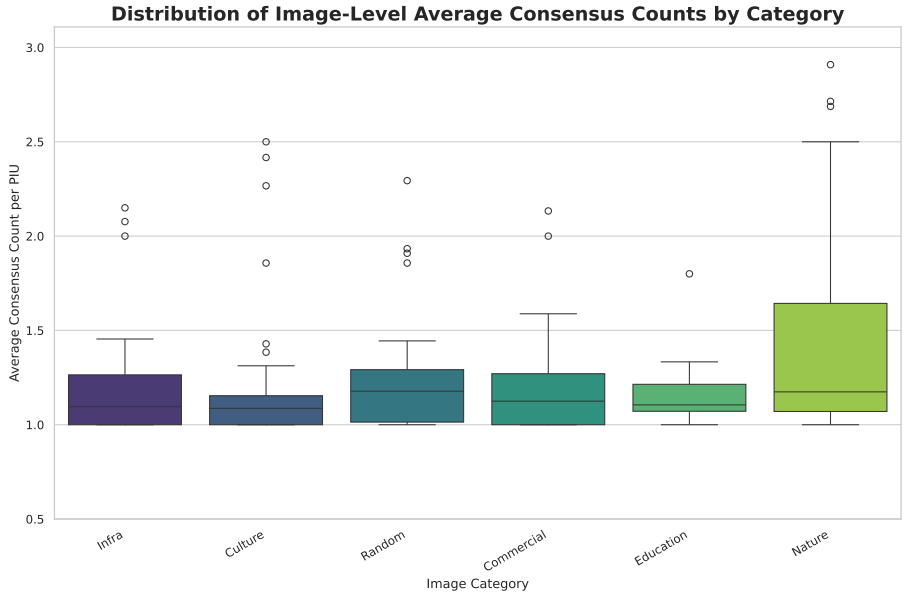
underscoring its lack of stability under these conditions.

A.5.2 HAR and MAR evaluations

Figure 11 illustrates an example of bidirectional evaluation highlighting differences between human- and model-generated captions. The human caption in this example includes narrative elements such as "they have been working for a few days" that are not directly grounded in the visual content. This reflects a broader trend in which humans often elaborate beyond visible cues, drawing on personal experience or cultural associations. By contrast, the model-generated caption (GPT-4o) introduces an error by incorrectly identifying a wooden plank in Person 2's hand, yet it provides a more exhaustive description of the observable scene. Such differences emphasize the asymmetry between human and model captions - humans enrich descriptions



(a)



(b)

Figure 10: Figure (a) Depicts a Boxplot vizualisation of the DCScore (HAR and MAR) of Random images with in length constrained caption setting across three models(Trends stay consistent) and Figure (b) shows category-wise Distribution of PIU consensus scores, showing limited descriptive overlap among human annotators

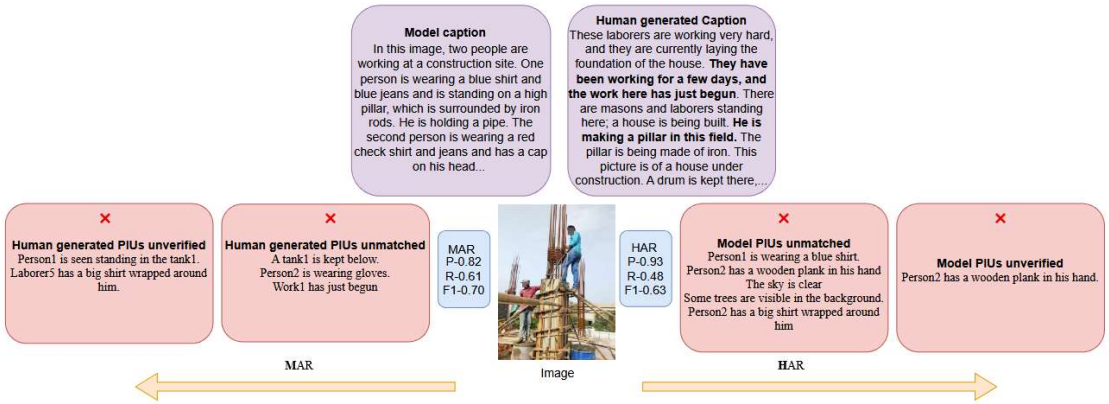
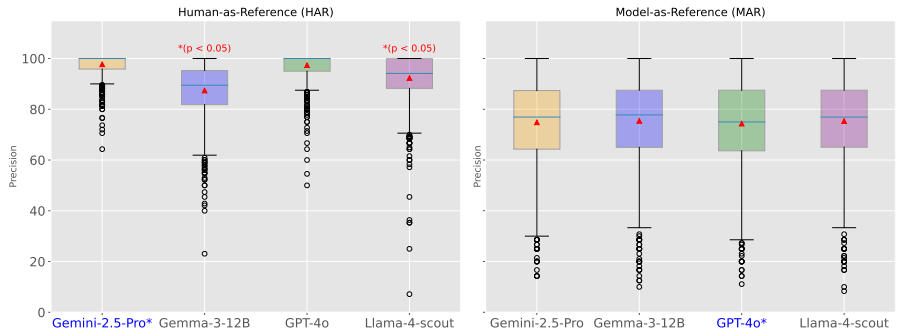


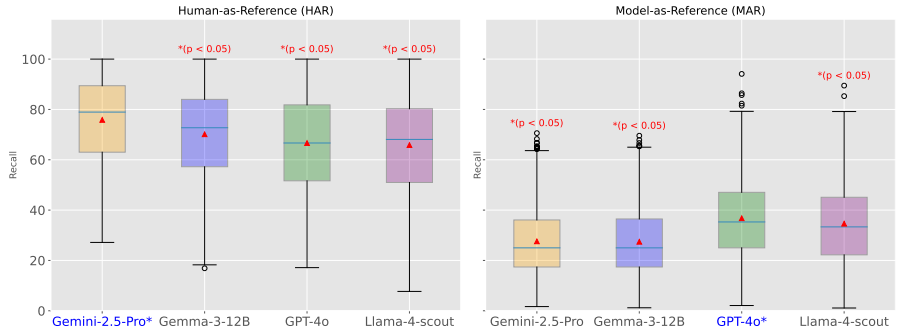
Figure 11: Example of bidirectional evaluation highlighting differences between human- and model-generated captions.

with context, while models prioritize coverage of the visible scene.

In addition to the F1 score results reported in the main text (Section 5.1), we provide a detailed



(a) HAR and MAR Precsion score across models



(b) HAR and MAR Recall score across models

Figure 12: Precison and Recall score for Random category across models

analysis of precision and recall for both Human-as-Reference (HAR) and Model-as-Reference (MAR) evaluations in Figures 12a and 12b, respectively. These figures show precision and recall distributions across models and highlight the trade-offs between exhaustive coverage (high recall) and alignment with reference content (high precision). The additional metrics help contextualize the DC-SCORE differences observed in the main text and illustrate model-specific tendencies toward exhaustive coverage and reduced selectivity.

Figure 16 shows the boxplots of F1 scores across five categories - Commercial, Infrastructure, Nature, Culture, Education - as well as Random, for the four evaluated models under both HAR and MAR settings.

A.5.3 HAR and MAR evaluations - English

To address the dependency and adaptability of the proposed bi-directional framework to languages, we conducted a full evaluation by translating the dataset and repeating the intermediary pipeline steps in English. We used an LLM-based clause-to-clause translation for human transcripts to preserve the original syntactic structure and 'spontaneous' nature of the speech. As shown in the Table 5, the primary rankings remain consistent.

Gemini-2.5-Pro remains the statistically significant leader in the HAR (Description) direction, while GPT-4o remains the leader in the MAR (Retrieval) direction. This confirms that our main paper's conclusions are not artifacts of the Hindi pipeline.

Gemini-2.5-Pro shows stability in HAR , indicating that its factual grounding capabilities are equally robust in both Hindi and English. In the MAR setting, GPT-4o sees a significant performance boost in English.

This might suggest that GPT-4o's English cap-

Table 5: HAR and MAR F1 scores across Hindi and English evaluations. Δ denotes the change from Hindi to English.

Model	Set.	Hindi	English	Δ
Gemini-2.5-Pro	HAR	**0.8263**	**0.8354**	**+0.0091**
	MAR	0.4212	0.3530	-0.0681
Gemma-3-12B	HAR	0.7536	0.7516	-0.0021
	MAR	0.4109	0.3651	-0.0458
GPT-4o	HAR	0.7593	0.6816	-0.0777
	MAR	**0.5032**	**0.6168**	**+0.1136**
Llama-4-scout	HAR	0.7430	0.7744	+0.0031
	MAR	0.4894	0.3923	-0.0970

Table 6: Representative hard samples from the Nature category illustrating dataset-specific granularity mismatches.

Dataset	Human Transcript	Model Caption
Places	भारत में कुछ मूर्ति–यां बनी हुई हैं, हरे–भरे पेड़ दिखाई दे रहे हैं *(Some statues are built in India; green trees are visible)*	यह तस्वीर माउंट रशमोर नेशनल मेमो–रियल का एक प्रमुख दृश्य दिखाती है *(A prominent view of Mount Rushmore National Memorial)*
VAANI	ये मैदान बहुत ज़्यादा खूबसूरत हैं, देखने में बहुत अच्छे लगते हैं *(These grounds are very beautiful and look pleasant)*	यह एक पार्क या खेल के मैदान का दिन का दृश्य है *(A daytime view of a park or playground)*

Table 7: Representative hard samples from the Education category highlighting inference and abstraction errors.

Dataset	Human Transcript	Model Caption
Places	एक कमरे में तीन महिलाएं और कई छोटे बच्चे बैठे हैं *(Three women and many small children are sitting in a room)*	यह तस्वीर एक स्कूल की कक्षा के अंदर की है *(This picture is from inside a school classroom)*
VAANI	यहाँ एक रूम है जि–सकी दीवार का रंग हल्का है और किता–बें रखी हैं *(A room is visible with light-colored walls and books)*	यह एक कमरे के अंदर का दृश्य है, जिसमें अलमारी और किताबें हैं *(An indoor room scene with a cupboard and books)*

tions align much more closely with human descriptive patterns than its Hindi captions. Crucially, the leaderboard remains unchanged, validating the robustness of our original analysis.

A.6 Category-wise Difficulty Discrepancy Across Datasets

While Nature and Educational categories exhibit a slightly higher proportion of *Easy* samples than Infrastructure in the Places Audio Captions dataset (Figure 15), this separation is substantially weaker than that observed for VAANI (Section 5.2.1). This suggests that category-level visual salience alone is insufficient to explain difficulty under bidirectional evaluation.

Our analysis suggests that difficulty is governed by the *granularity mismatch* between image content and reference transcripts. In the Places audio captions dataset, human references are often sparse or under-specified, leading to precision penalties when models introduce visually valid but textually unsupported details like the Nature category examples shown (Table 6, Places). In contrast, VAANI transcripts are dense and experiential, resulting in recall penalties when models omit subjective or affective predicates present in human speech (Table 6, VAANI). A similar pattern holds for the Education category (Table 7), where Places failures are driven by functional inference, while VAANI failures arise from abstraction over fine-grained enumerations.

A.6.1 Proposed Bi-directional Framework vs Existing Metrics

To determine if existing metrics could suffice, we computed BLEU-4, METEOR, ROUGE-L, and BERTScore for our dataset and analyzed their Spearman correlation (ρ) with the human judgments collected in Section 5.3.

Table 8: Comparison of proposed bidirectional framework vs. existing metrics such as BLEU-4, METEOR, ROUGE-L, and BERTScore

Metric	ρ - MAR	ρ - HAR
Bidirectional-HAR/MAR (Our proposed framework)	**0.263**	**0.121**
BERTScore	0.206	-0.065
METEOR	0.165	-0.058
ROUGE-L	0.156	-0.002
BLEU-4	-0.075	-0.075

As shown in the Table 8, our proposed metric (Bidirectional-HAR/MAR) is the only method that aligns closest positively with human judgment in both directions. Relying on the other metrics mentioned would be statistically misaligned with the human judgments recorded, necessitating the proposed bidirectional framework.

1. In the Human-as-Reference setting, all tradi-

tional metrics exhibit negative correlations (e.g., BLEU-4 at -0.075, BERTScore at -0.066). This confirms that lexical and embedding-based matching fail when evaluating spontaneous, speech . They penalize the valid but structurally colloquial human transcripts that our PIU-based approach correctly handles.

2. In the Model-as-Reference setting, our MAR score (0.264) still outperforms the BERTScore (0.206). This indicates that evaluating human input requires the granular, fact-based decomposition offered by our pipeline, rather than the broad semantic similarity measured by embeddings.

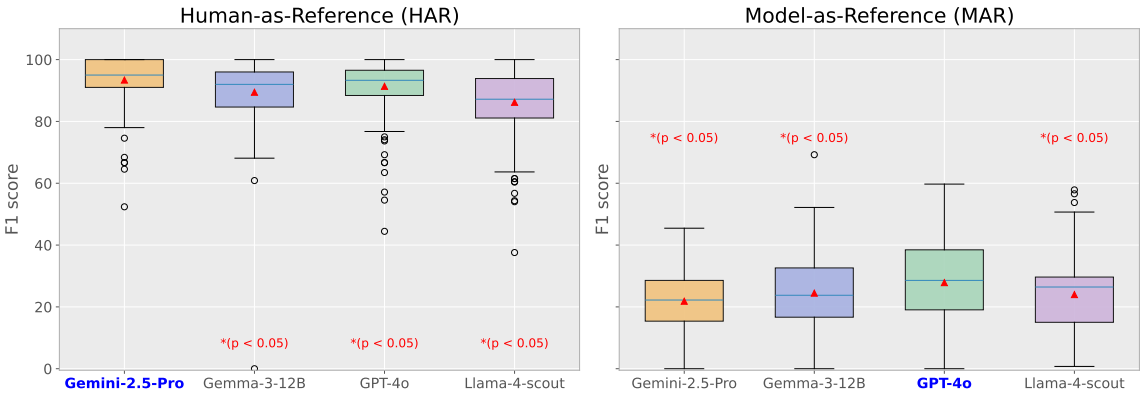


Figure 13: HAR and MAR F1 scores for Random subset of Places Audio Captions dataset across Models

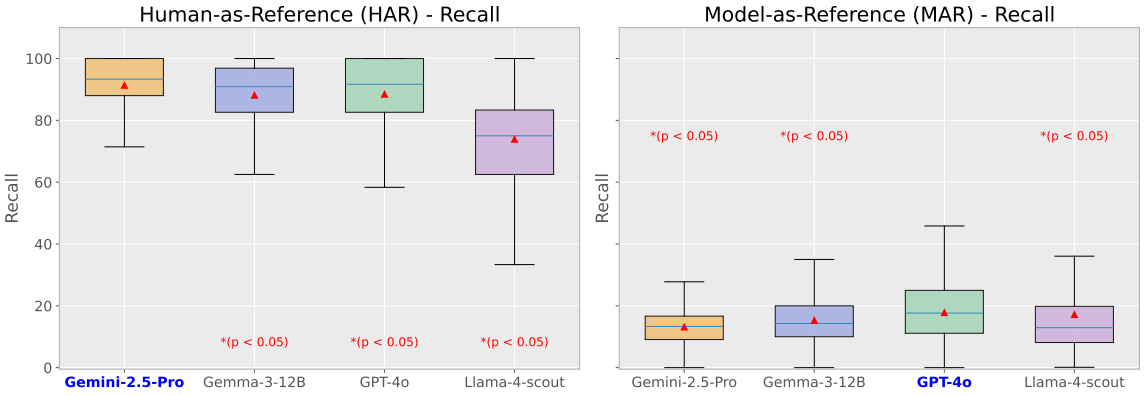


Figure 14: HAR and MAR Recall scores for Random subset of Places Audio Captions dataset across Models

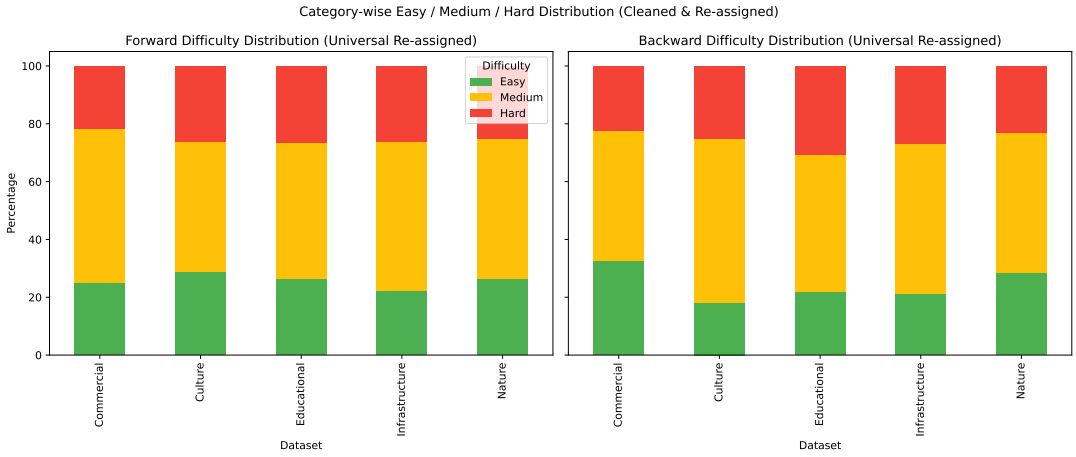


Figure 15: Difficulty distribution of images across categories in both HAR/MAR (F1) setting. nature and education categories contain the highest proportion of 'Easy' samples in HAR setting.

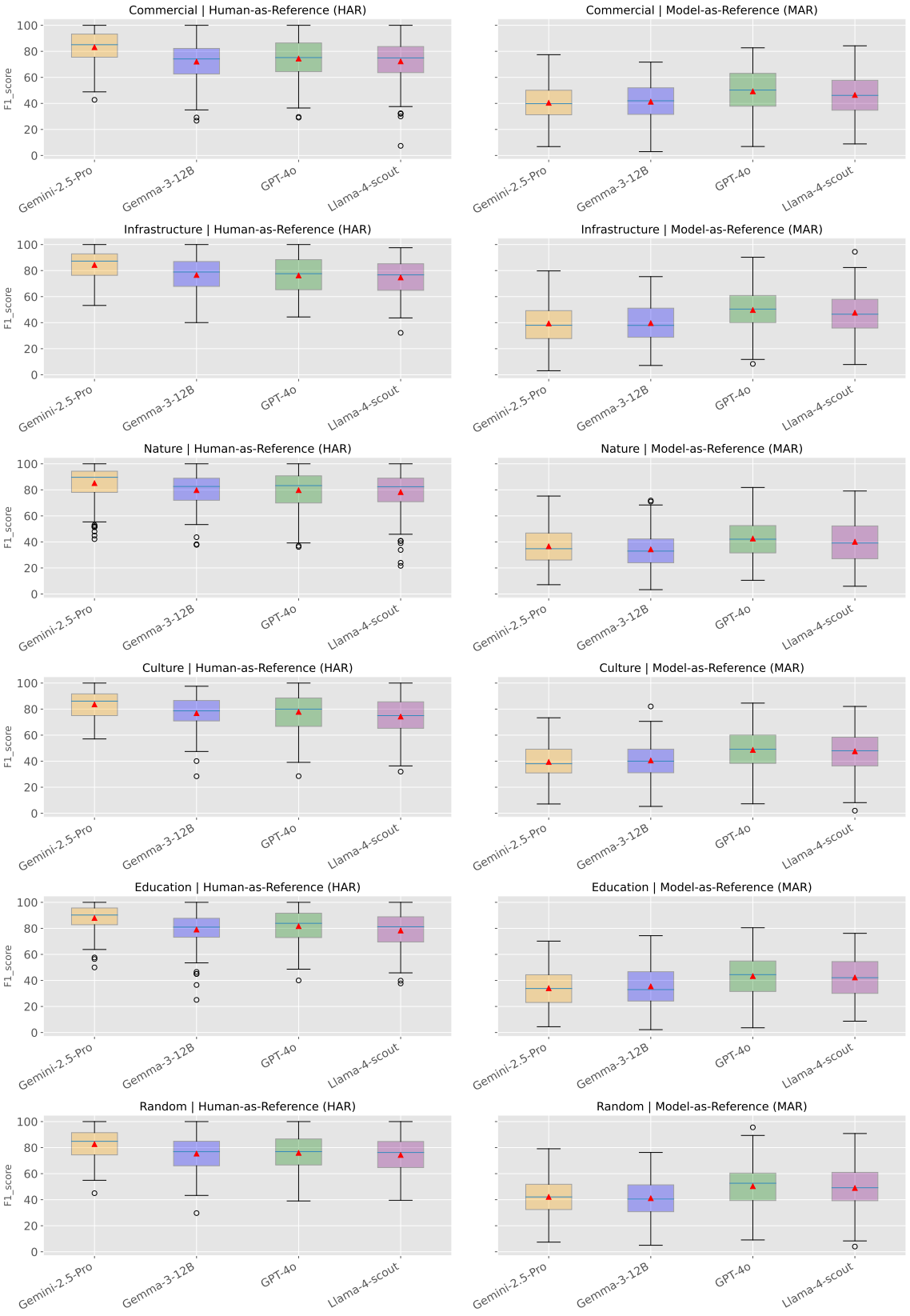


Figure 16: F1 score across categories - Commercial, Infrastructure, Nature, Culture, Education and Random for the four MLLMs under HAR and MAR settings.

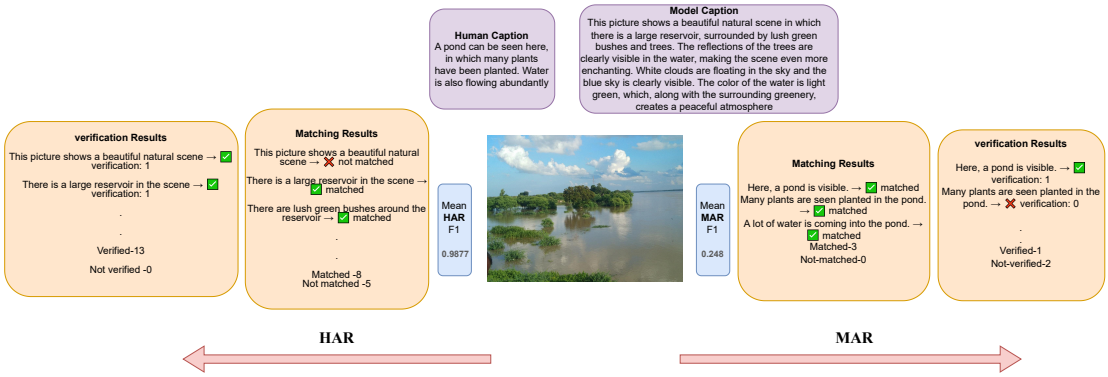


Figure 17: Example of an image in Nature, where it is classified as 'Easy' in HAR but 'Hard' in MAR.

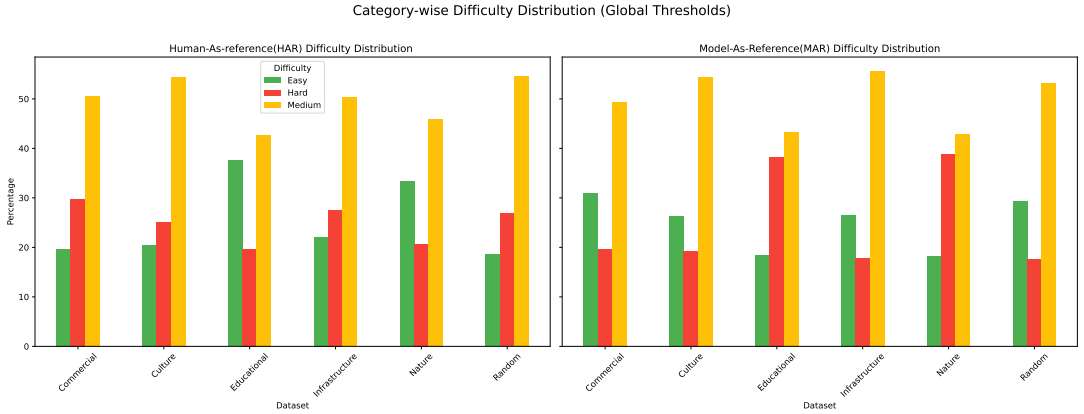


Figure 18: Difficulty distribution of images across categories in both HAR/MAR (F1) setting. Nature and Education categories contain the highest proportion of 'Easy' samples in HAR setting.

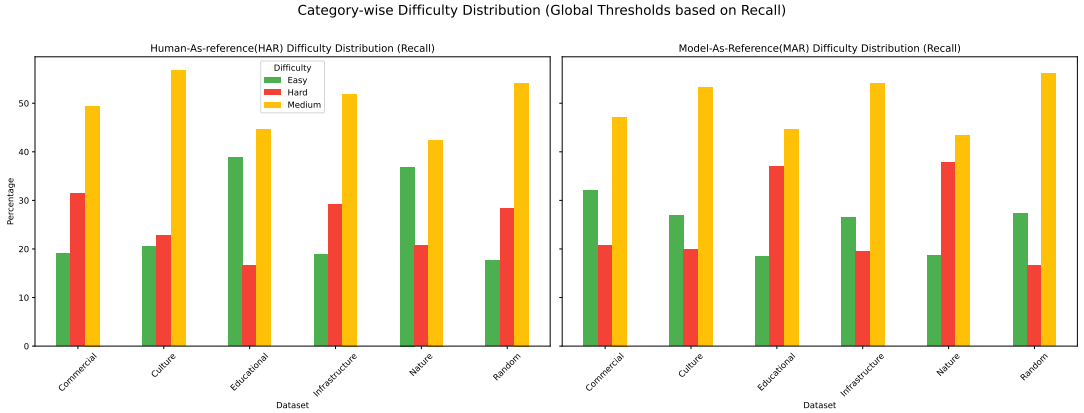
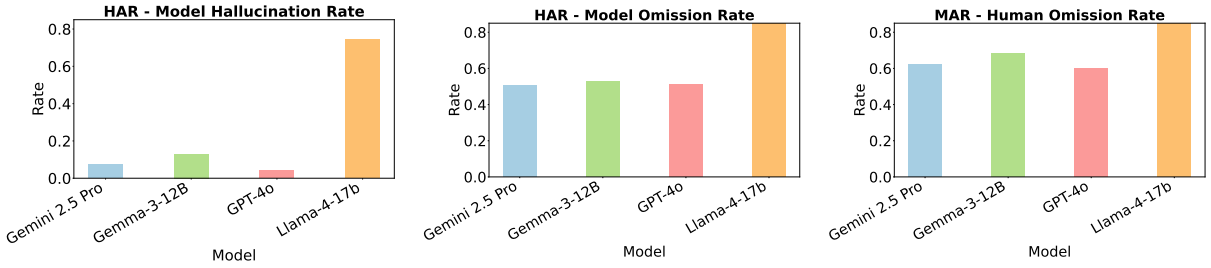
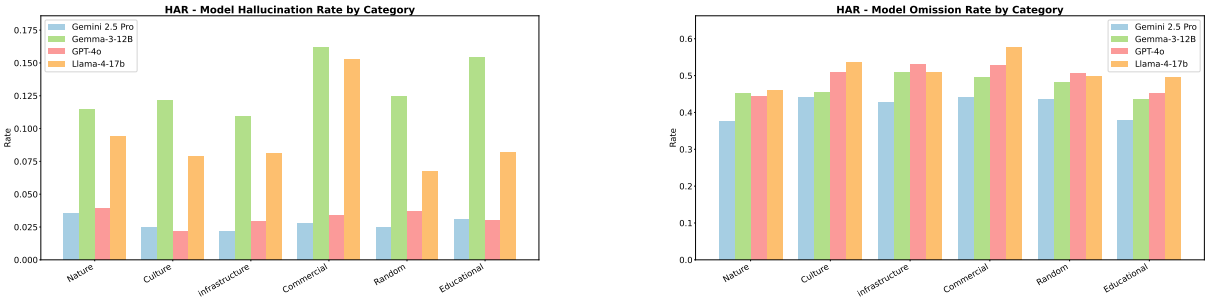


Figure 19: Difficulty distribution of images across categories in both HAR/MAR (Recall) settings. Nature and Education categories contain the highest proportion of 'Easy' samples in HAR setting.

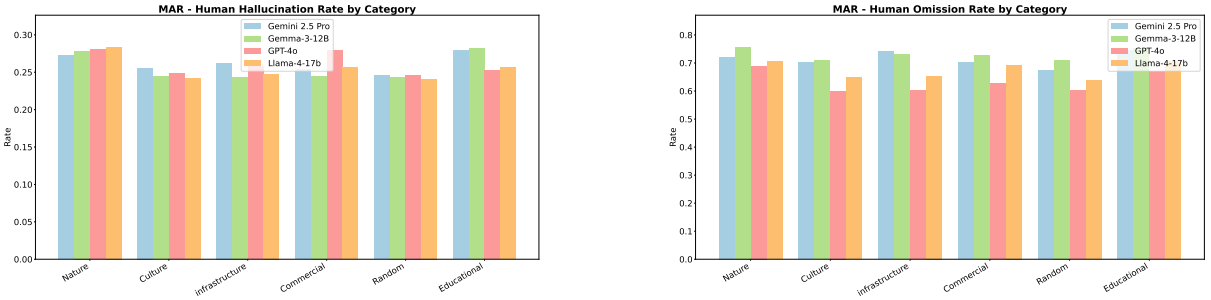


(a) Model Hallucinations (b) Model Omissions (c) Human Omissions

Figure 20: Bidirectional hallucination and omission rates for word-length constrained model captions: (a) Model hallucinations in the HAR setting, (b) Model omissions in the HAR setting, and (c) Human omissions when model captions serve as reference (MAR)



(a)



(b)

Figure 21: Figure (a) Depicts Hallucination and Omission rates across categories in HAR setting and (b) Hallucination and Omission rates across categories in MAR setting, with Gemma showing the highest hallucination across categories.

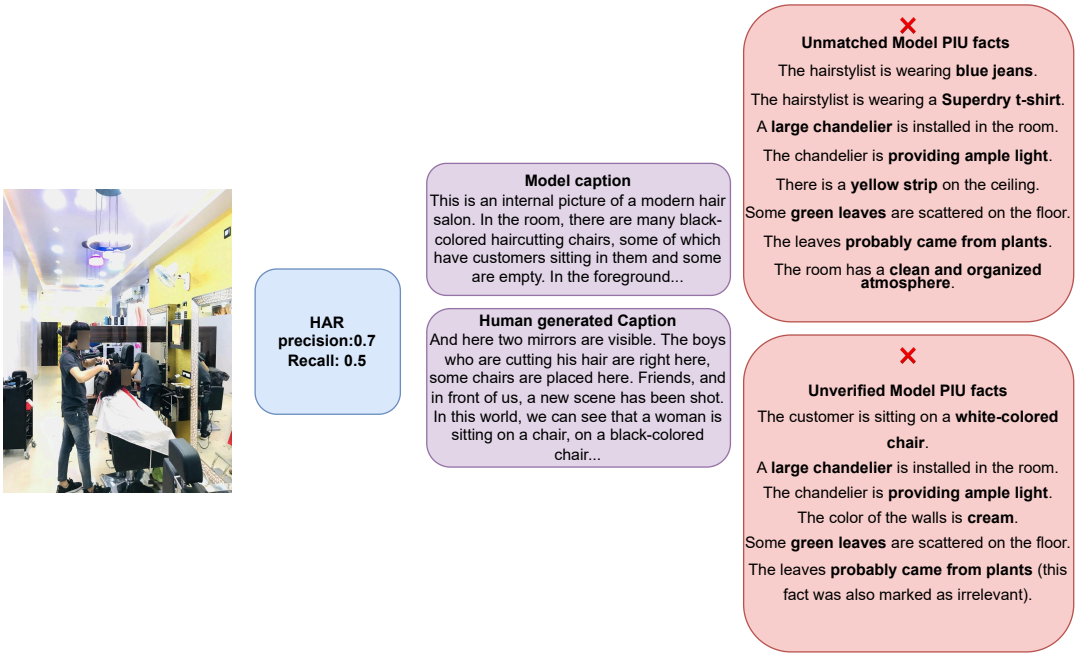


Figure 22: Example of a Commercial category image for which **Gemma** generated caption with Low HAR precision and recall. In contrast, the Gemini model generated caption boasts precision of 0.96 and Recall of 0.68.