

Generating Vehicular Icon Descriptions and Indications Using Large Vision-Language Models

James Fletcher¹, Nicholas Dehnen¹, Seyed Nima Tayarani Bathaie¹, Aijun An¹,
Heidar Davoudi², Ron Di Carantonio³, Gary Farmaner³

¹York University, Toronto, Canada, ²Ontario Tech University, Oshawa, Canada,

³iNAGO Co, Toronto, Canada

{jfletche, ndehnen, nimatb, aan}@yorku.ca, heidar.davoudi@ontariotechu.ca,
{rond, garyf}@inago.com

Abstract

To enhance a question-answering system for automotive drivers, we tackle the problem of automatic generation of icon image descriptions. The descriptions can match the driver's query about the icon appearing on the dashboard and tell the driver what is happening so that they may take an appropriate action. We use three state-of-the-art large vision-language models to generate both visual and functional descriptions based on the icon image and its context information in the car manual. Both zero-shot and few-shot prompts are used. We create a dataset containing over 400 icons with their ground-truth descriptions and use it to evaluate model-generated descriptions across several performance metrics. Our evaluation shows that two of these models (GPT-4o and Claude 3.5) performed well on this task, while the third model (LLaVA) performs poorly.

1 Introduction

Vehicle dashboard icons convey critical information to drivers, who must quickly understand the symbols' meaning and take appropriate action. However, many drivers are unfamiliar with these icons, emphasizing the pressing need for a virtual assistant that can explain the icons' meanings. For example, when presented with a dashboard icon resembling a steaming cup, the driver might naturally ask "What does that cup icon mean?" The correct response is that this is a warning from the vehicle's driver attention system.

The iNAGO netpeople® Assistant is a proprietary voice-based virtual assistant platform for automotive drivers. It can answer drivers' questions based on knowledge extracted from text documents, such as car manuals. However, netpeople currently struggles with icon-related inquiries because its text-based knowledge base lacks icon descriptions. This gap means driver's questions about icons cannot be matched to any existing knowledge items.

Currently, no conversational system for drivers can answer questions about dashboard icons.

To address this, we aim to automatically generate text descriptions for icon images, enabling netpeople to include questions and answers (QAs) about dashboard icons in its knowledge base. This task presents several challenges. First, existing image description systems are trained mainly on natural images, whereas icon images are drawings. Second, understanding an icon's function, beyond its visual description, requires context from the manual and is harder than typical image captioning. Training and evaluating a model that generates both visual and functional icon descriptions necessitates a labeled dataset, which currently does not exist. Third, while many metrics for text generation are available, identifying the most suitable metrics for evaluating both visual and functional descriptions of dashboard icons is crucial.

In this work, we compile a dataset of 408 vehicle dashboard icon images and their corresponding names/functions, which we collected from 42 vehicle manuals. We use state-of-the-art multimodal Large Vision-Language Models (LVLMs) (i.e., GPT-4o (OpenAI et al., 2024), LLaVA-NEXT (Liu et al., 2023) and Claude 3.5 (Anthropic PBC, 2024)) to generate natural English descriptions of each icon's visual design and function. Such descriptions can form QA pairs for netpeople's knowledge base. We assess model performance using standard performance metrics and human evaluation. The key contributions of this work are:

- We create a new image description dataset with human-generated visual and functional descriptions for vehicle dashboard icons ¹.
- Using this new dataset, we show that several state-of-the-art and generically trained

¹Available: <https://github.com/yorku-datamining-lab/generating-vehicular-icon-descriptions>

LVLMs can perform well on this icon description task.

- We compare several automatic performance metrics against human evaluation scores and found that SBERT cosine similarity scores are most consistent with human evaluation scores.

2 Related Work

Efforts to verbalize images through image captioning (Chan et al., 2023) and summarization (Celis and Keswani, 2020) use retrieval (Lindh et al., 2018) or generation methods (Vinyals et al., 2015). Recently, researchers have combined these by retrieving image-caption pairs and inputting them into generation models (Ramos et al., 2023).

LVLMs use Large Language Models (LLMs) in vision-language tasks either as schedulers (Chen et al., 2022; Surís et al., 2023), where the LLM manages various visual models as plug-and-play modules based on specific task requirements, or as decoders (Zhu et al., 2024), enabling cross-modal knowledge transfer. With the increasing need for larger language model backends, approaches like InstructBLIP (Zhu et al., 2024) and LLaVA (Liu et al., 2023) collect extensive human instruction datasets to train larger LVLMs. These models then undergo end-to-end training, enhancing the LLMs with visual reasoning capabilities. GPT-4o (OpenAI et al., 2024) and Claude 3.5 (Anthropic PBC, 2024) are advanced multimodal models that process text and image inputs to generate text outputs, exhibiting human-level performance on various professional and academic benchmarks.

In this work, we utilize three recent LVLMs, i.e., GPT-4o, Claude 3.5, and LLaVA-NEXT, to generate descriptions of dashboard indicator icons. Our goal is to enable a QA system to answer drivers' questions about these icons. To the best of our knowledge, this is the first application of AI models for generating such descriptions.

3 Methodology

3.1 Dataset Creation

We collected images of dashboard icons from 42 vehicle manuals from four different manufacturers (see Appendix A for details). All the manuals were available on the internet. To facilitate the generation of functional descriptions of icons, we considered only the manuals available in HTML files, which enabled automated extraction of icon

images along with the surrounding context text. This was achieved by identifying each image tag in the main body of the document, ascending 2-3 levels up the HTML parse tree from the image tag, and selecting all of the text under that parent node. This creates the input part of each example in the dataset. Table 1 shows an example in our dataset.

While dashboard icons' visual designs are standardised (ISO, 2021), manufacturers often make minor embellishments. Hence, to remove exact image duplicates but preserve different image variants, we computed a 64-bit dHash image hash for each icon based on the horizontal gradient of a down-sampled versions of each original icon image (Buchner, 2024). After removing images with duplicate hashes, 408 unique icon images remained.

Two separate ground-truth descriptions of icon each image were produced: (1) a visual description of the image, focusing on the recognizable components that could be seen within the image; and (2) a functional description that described the purpose and intent of the icon, based on the appropriate manual text. The importance of separating these two types of descriptions is that the visual and functional descriptions form the question and answer, respectively, in the knowledge base of netpeople, which we are seeking to enhance.

To create the ground-truth functional description for each example, two native English speakers read the relevant part of the car manual for each icon and extracted or created its functional description. Each icon has a single ground-truth functional description since each icon has one specific indication. In contrast, for creating the ground-truth visual descriptions of the icons, 28 fluent English-speaking human annotators are used to collect diverse visual descriptions for an icon, as different people may describe an image differently. For example, considering the different visual descriptions of the cup icon shown in Table 1.

To collect the visual descriptions of the icons, we created a web interface. The starting page provides instructions and examples to the annotators, followed by subsequent pages where each page displays an icon image. On these pages, annotators can enter their descriptions and indicate the degree of difficulty in describing the image on a scale of 1-5, with 5 meaning the most difficult. The difficulty labels will be used in the analysis when evaluating the description-generation models. Appendix A shows a snapshot of the instruction page and an example page displaying an icon and collecting the


Input	Ground Truth Descriptions
	<p><i>Visual:</i> "This amber dashboard icon depicts a cup and saucer. Three wavy lines above the cup represent the idea that the cup contains a hot drink." "The pictogram depicts an orange coloured cup placed on a saucer. The steam is coming out of the cup." "The image shows an amber coffee mug on a coaster. Wavy vertical lines indicate steam rising from the coffee mug."</p>
See Driver Condition Monitor (Amber)	<p><i>Functional:</i> "The icon indicates that the vehicle's driver condition monitor system has detected that the driver is presenting signs of high fatigue levels."</p>

Table 1: A single example from our dataset: Driver Condition Monitor

annotator's inputs.

As a result, a total of 408 examples were created. A subset of 20 examples was randomly selected for use in few-shot prompting, and the remaining 388 samples formed the test set for evaluating models.

3.2 Automatic Generation of Image Descriptions

Given an icon image and its context description from a car manual, the task is to generate both a visual description and a functional description of the icon. The visual description should explain what the icon looks like, while the functional description should explain what the icon indicates (e.g. see the context and ground truth descriptions for the cup icon shown in Table 1).

Three pre-trained state-of-the-art LVLMs were used in this study: GPT-4o (OpenAI et al., 2024), LLaVA-NEXT:34b (Liu et al., 2023) and Claude 3.5 (Anthropic PBC, 2024). The steps for using these models for automated icon description generation were straightforward: A base64-encoded icon and corresponding context were supplied to the models along with an appropriate prompt (see Appendix B), and the model output was collected; containing both, visual and functional descriptions separately.

These models were pre-trained for general-purpose tasks. To alleviate hallucination, we experimented with few-shot prompts in addition to zero-shot prompts. To select the few shot examples, we choose the k icons from the training set that were closest to the query icon by computing the Hamming distance between the image hash of the query image and the image hash of each of the training images. This was made possible by the properties of the dHash image hashing method, where similar input images produce similar output hashes (Buchner, 2024).

The following is a prompt for zero-shot: "You are an AI visual assistant specialized in interpreting icons displayed on the dashboard of a vehicle.

An icon communicates important information about the vehicle to the driver. You are seeing an image of a single dashboard icon. Briefly describe the dashboard icon depicted in the image, focusing on the visual content of the image and meaning of the icon. Limit your response to 2 sentences. The first sentence should describe the visual content. The second sentence should describe the icon's meaning. Format your response as a JSON object with the following keys: 'visual_content', 'meaning'. The image has the following associated text: ..."

Appendix B shows prompts for k -shot (for $k = 1, 3, 5$). Of the three models, we found that LLaVA required the most explicit prompting in order to produce acceptable output. To make a fair comparison, we selected the best prompt for LLaVA and used the same prompt for all three models. We also found it difficult to prevent the models from including functional descriptions, even when explicitly prompted to only provide visual descriptions. This is why we prompted the models to generate both visual and functional descriptions together and then separate them in a JSON object.

4 Evaluation

In our evaluation, we primarily considered the performance of different LVLMs, effectiveness of few-shot prompting, and correlation of automated performance metrics with human evaluation. We also evaluated the impact of description type and input type on model performance.

4.1 Automatic Metrics

A variety of automatic metrics were used to evaluate the model-generated icon descriptions against the human-generated ground truth. Traditional rule-based metrics such as ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and GoogleBLEU4 (Wu et al., 2016) (a variant of the conventional BLEU score) were used, along with several newer embedding-based metrics such as BERTScore (Zhang et al., 2020) and CLIP-Score and

Model	k-shot	Visual Description							Functional Description						
		BS	SB	CL	rCL	GB4	M	R	BS	SB	CL	rCL	GB4	M	R
Claude 3.5	0	0.73	0.68	0.79	0.81	0.16	0.24	0.39	0.79	0.86	0.75	0.82	0.21	0.33	0.47
	1	0.74	0.70	0.79	0.81	0.16	0.25	0.40	0.80	0.84	0.74	0.82	0.22	0.33	0.49
	3	0.76	0.70	0.79	0.82	0.17	0.27	0.41	0.81	0.85	0.74	0.82	0.24	0.34	0.52
	5	0.78	0.70	0.78	0.83	0.21	0.29	0.45	0.82	0.86	0.73	0.82	0.27	0.35	0.54
GPT-4	0	0.76	0.70	0.77	0.81	0.18	0.26	0.42	0.82	0.85	0.73	0.81	0.28	0.32	0.53
	1	0.81	0.72	0.79	0.84	0.25	0.32	0.53	0.84	0.88	0.72	0.81	0.32	0.35	0.58
	3	0.82	0.73	0.79	0.85	0.30	0.34	0.57	0.85	0.88	0.71	0.81	0.34	0.36	0.61
	5	0.83	0.73	0.79	0.85	0.33	0.36	0.58	0.85	0.89	0.71	0.81	0.35	0.38	0.62
LLaVA	0	0.70	0.61	0.75	0.78	0.11	0.19	0.31	0.74	0.77	0.72	0.79	0.14	0.26	0.35
	1	0.73	0.62	0.75	0.79	0.14	0.14	0.28	0.79	0.79	0.71	0.79	0.20	0.27	0.43
	3	0.72	0.61	0.74	0.78	0.13	0.20	0.34	0.78	0.78	0.70	0.79	0.19	0.27	0.43
	5	0.72	0.62	0.75	0.78	0.13	0.21	0.35	0.78	0.78	0.71	0.79	0.19	0.28	0.43

Table 2: Results for all metrics for the k-shot prompting evaluation with one input level (image-and-context) and four prompting levels (k=0, 1, 3, 5). Metrics used: BERT-Score (BS), SBERT-Score(SB), CLIP-Score (CL), RefCLIP-Score (rCL), Google-BLEU4 (GB4), METEOR (M), ROUGE (R).

RefCLIP-Score (Hessel et al., 2021). In addition, we also used SBERT (Reimers and Gurevych, 2019) to compute the embeddings of the generated descriptions and ground-truths, which we compared using the cosine similarity score (which we refer to as SBERT-Score). While the various scores operate on different principles, in all cases higher values represent closer agreement between model outputs and ground truth.

4.2 Human Evaluation

We additionally conducted a human evaluation study comparing the three models (Claude 3.5, GPT-4o, and LLaVA) on their ability to generate visual descriptions for 60 images (15% of the entire dataset). The images were randomly selected from the most dissimilar images in the test set based on the Hamming distance ($d_{\min} = 21$) of their dHash (see Section 3.1). Six participants each rated 30 descriptions (generated by the models with 3-shot prompts) on a one to five Likert scale. A balanced incomplete block design (see Appendix C.6.1) was chosen to minimize order effects, and each participant was assigned to two out of four 15 image blocks, resulting in three ratings per description and thus $3 \times 60 = 180$ ratings in total.

We opted not to have human evaluators assess the generated functional descriptions of icons due to the potential lack of knowledge about each icon’s functional indications. Providing car manuals or ground-truth functional descriptions to address this knowledge gap could have inadvertently influenced the evaluation of visual descriptions, creating a confounding factor in the experiment. We wanted evaluators to judge visual descriptions based solely on

images, allowing for a range of valid interpretations beyond a single ground truth. By withholding functional information, we maintained consistency in our evaluation and prevented potential bias, ensuring that the ratings of visual descriptions remained uninfluenced by functional details. This approach allowed us to focus on obtaining unbiased evaluations of the visual aspects while acknowledging the limitations in assessing functional descriptions without appropriate domain knowledge.

4.3 Results and Findings

We show the evaluation results and address several research questions.

RQ1: Which model performs best on this task?

Table 2 compares the three models with k -shot prompts on all the performance metrics. GPT-4o generally performed best for both description types, with Claude 3.5 close behind, and LLaVA performing relatively poorly. Metrics like BERT-Score, SBERT-Score and Google-BLEU4 produced similar rankings, while Meteor and Rouge rankings were also similar. CLIP and RefCLIP aligned with others for visual descriptions, but quite different rankings for functional descriptions. This is likely because CLIP has no access to the context text associated with an image when it generates the reference description, to which it then compares the generated functional descriptions. CLIP is forced to generate (and potentially hallucinate) a reference functional description purely from the image itself. To further analyse the results, we use SBERT-Score because it aligns best with human judgement scores to be presented later in this section. The



Icon	Model	Generated Visual Description	SBERT Score	Human Eval.
	GPT-4	This amber dashboard icon depicts a cup of steaming hot beverage, such as coffee or tea.	0.70	3.7
	LLaVA	The icon depicts a stylized representation of a cup with steam rising from it.	0.69	4.0
	GPT-4	This dashboard icon depicts a vehicle headlight with five horizontal lines extending to the left, indicating the light beams.	0.70	3.7
	LLaVA	The icon depicts a headlight with a snowflake inside, representing icy road conditions while the high beam is on.	0.44	1.0

Table 3: Examples of visual descriptions generated by GPT-4 and LLaVA with 3-shot prompting for 2 icons.

grand mean SBERT-Score across all models was 0.77 ± 0.13 . GPT-4o achieved the highest similarity of 0.8 ± 0.12 , which is 4.46% above average². Claude 3.5 was just slightly worse than GPT-4o at a mean of 0.77 ± 0.12 , but still 1.06% higher than average. Conversely, LLaVA demonstrated a mean cosine similarity of 0.7 ± 0.14 , which is 8.65% below average. While the difference between Claude 3.5 and GPT-4o was small, LLaVA came in far behind the two models, scoring much worse on average. A Friedman test showed that all differences were statistically significant ($p < 0.001$, see Appendix C.1). For example, consider the two icons in Table 3 with visual descriptions by GPT-4o and LLaVA. Both correctly describe the first icon, but LLaVA’s description of the second, more challenging icon is incorrect, mentioning a non-existent snowflake. Claude 3.5 performed similarly to GPT-4o. These examples show that LLaVA can match other models when it detects visual content correctly, but it often produces hallucinations when it fails. Despite many ‘vision failures,’ LLaVA’s scores were only slightly lower due to automated metrics focusing on word matches rather than meaning differences, which will be discussed further.

Finding: GPT-4o performed best on the task, but is followed closely by Claude 3.5. LLaVA performed significantly worse.

RQ2: Does few-shot prompting improve model performance?

For GPT-4o and Claude 3.5, the metrics show the performance generally improves with increasing k in the few-shot prompting. However, for LLaVA, whether few-shot is better than zero-shot depends on the metrics and the type of description. For visual descriptions, 1-shot is better than zero-shot for most metrics except Meteor and Rouge. For

²Percentages calculated as $\frac{\text{SBERT}_{\text{score}} - \text{mean}}{\text{mean}}$ and reported to two decimal places.

functional descriptions, 1-shot is better than zero-shot on all metrics except CLIP. It is interesting to see that when $k > 1$, the performance of k -shot decreases for LLaVA. To see whether the improvement is significant, we conducted a Friedman test (see Appendix C.2), which shows that k -shot has statistically significant improvements in SBERT-Score over the 0-shot baseline for both GPT-4o and Claude 3.5. Claude 3.5 showed the largest improvement at $k = 5$ (+1.3%)³. GPT-4o had the highest gains at $k = 5$ (+3.73%), with $k = 3$ (+3.34%) and $k = 1$ (+3.03%) close behind. All improvement for $k > 1$ in GPT-4o were significant, although differences for $k \in [1, 3, 5]$ were minor. LLaVA showed no significant improvements for higher k levels. In few-shot prompting, prompts were selected from the training based on image similarity (see Section 3.2). An ablation study using GPT-4o alone compared this method to random selection, finding minimal improvement in visual descriptions (+0.27%) and a slight decrease in functional descriptions (−0.23%) at all k levels. These findings suggest that LVLMS generally benefit from few-shot prompting, though the impact varies. Claude 3.5 needed five examples for significant improvements, GPT-4o just one. However, for GPT-4o, using more than three examples may not justify the token cost. Few-shot prompting primarily results in style transfer of ground truth writing style, but does not improve the vision component (see Appendix D). Thus, a misinterpreted image may be described in a style similar to the ground truth, but the semantics will still differ.

Finding: 5-shot prompting improves GPT-4o and Claude 3.5 performance most, while LLaVA shows no benefit from few-shot prompting.

RQ3: To what extent does description type affect scores?

We observed significant differences in the scores between functional and visual descriptions com-

pared to the ground truth. The average SBERT-Score on functional description was 0.85 ± 0.1 , while the mean visual description SBERT-Score was 0.69 ± 0.1 , a relative difference of 23.65% across all models. GPT-4o achieved both the highest average scores and the lowest gap between the two types at 22.36%. Claude 3.5 came second with a relative difference of 22.91%, whereas LLaVA placed last with a large 27.31% margin between mean scores for the two description types. These differences were statistically significant between all pairs of model and description type, as revealed by a Friedman test (see Appendix C.3). These differences may be attributed to several factors: The context from vehicle manuals likely plays a crucial role in enhancing the models’ understanding of an icon’s function. Conversely, the lower performance in visual descriptions highlights the challenges LVLMs face in interpreting complex graphical elements. While the models’ abilities of interpreting difficult icons is analyzed in Appendix C.5, this discrepancy could also be attributed to the nature of the ground truth, which was generated by humans. Depending on the annotator, descriptions might incorporate deeper domain knowledge and cultural understanding on the one hand, or resort to a basic description of geometric features and similarity with other known symbols on the other. More advanced models like GPT-4o may bridge this gap better due to their larger size and improved integration of visual and contextual understanding.

Finding: The image descriptions of all models score significantly worse than their functional counterparts. This may be influenced by the context, the vision capabilities, and the large variability in ways of describing images.

RQ4: Can models achieve comparable performance using only text, or only images?

Table 4 compares the three models for zero-shot prompting across three input levels (image and context, image only, context only) using SBERT-Score and METEOR metrics. All models performed best with both image and context. For visual descriptions, performance was worst without images. For functional descriptions, performance was worst without context.

Providing image and context generally performed best (SBERT-Score: 0.75 ± 0.13), while image- (0.7 ± 0.13) and context-only (0.68 ± 0.17) were less effective. For functional descriptions,

context-only performed almost as well as image and context (0.79 ± 0.13 vs. 0.83 ± 0.1). For visual descriptions, image-only was close to image and context (0.65 ± 0.11 vs. 0.67 ± 0.10). A Friedman test showed the differences for image & context were always statistically significant for both description types (see Appendix C.4). The only exception was GPT-4o, where image-only and image plus context did not significantly differ for visual descriptions. This experiment was conducted for $k = 0$, and similar behavior is expected for $k > 0$. These results highlight the benefit of multi-modal inputs, especially for visual tasks.

Model	Input	Visual		Functional	
		SB	M	SB	M
Claude 3.5	i + c	0.68	0.24	0.86	0.33
	i	0.66	0.24	0.75	0.25
	c	0.61	0.19	0.82	0.32
GPT-4	i + c	0.70	0.26	0.85	0.32
	i	0.70	0.25	0.80	0.26
	c	0.59	0.18	0.83	0.32
LLaVA	i + c	0.61	0.19	0.77	0.26
	i	0.60	0.19	0.67	0.17
	c	0.50	0.13	0.72	0.24

Table 4: Results for SBERT-score (SB) and METEOR (M) on zero-shot results with 3 input levels: image-and-context (i+c), image-only (i) and context-only (c).

Finding: Models generally performed best when given both context & images. Depending on the description type, using solely images or context resulted only in a small score difference.

RQ5: How do automated scores relate to human judgment?

The grand mean of all human ratings on the visual descriptions was 3.14 ± 1.35 . Individual means were 3.57 ± 1 for GPT-4o, 3.65 ± 1.17 for Claude 3.5, and 1.89 ± 1.18 for LLaVA. A mixed-effects model analysis found no significant difference between GPT-4o and Claude 3.5, but both significantly outperformed LLaVA. LLaVA overall scored very low (IQR: 1 – 2), suggesting that while LLaVA’s descriptions share some semantic similarity with ground truth, they lack or misrepresent crucial elements that human raters deem important. Inter-rater agreement analysis revealed strong consensus among participants (see Appendix C.6.3). The intraclass correlation coefficient (ICC) values indicate excellent agreement (Koo and Li, 2016), with $ICC(3, k) = 0.987$

(95% CI: [0.95, 1.0]). All participants consistently ranked the models in the same order: Claude 3.5 slightly outperforming GPT-4o, with LLaVA receiving notably lower scores. Appendix C.6.2 provides more detailed analysis on the correlation of automated metrics with human ratings, revealing that all correlations with the automated metrics were weak. Our analysis found SBERT cosine similarity most consistent with human judgments while providing easily interpretable scores. Traditional metrics like METEOR and ROUGE showed high correlation with human ratings but produced lower average scores with high standard deviations.

Finding: The human evaluation confirms that GPT-4o and Claude 3.5 are significantly better than LLaVA, with strong inter-rater agreement. Among the automatic metrics, SBERT-Score is most consistent with human ratings.

4.4 Generalizability of Findings

While our findings provide valuable insights into the current performance of LVLMs on vehicle icon description tasks, the specific performance gaps we observed between models may change as LVLMs continue to evolve rapidly. Nevertheless, our general findings - such as the importance of multi-modal inputs and the challenges in visual interpretation of abstract symbols - remain relevant. Future LVLM versions may address some of the current limitations, particularly in visual hallucinations and abstract symbol interpretation. Researchers applying these findings to new models should consider the specific architecture and training data of the models, as these factors significantly influence performance on specialized tasks like icon interpretation. Moreover, as automotive technology advances, the nature and complexity of dashboard icons may change, potentially requiring future reassessments of LVLM performance in this domain.

5 Conclusion

We have presented a novel application of large vision-language models to generation of vehicle dashboard icon descriptions. Our contributions include a novel task for automatic generation of visual and functional descriptions of automotive icons, enabling QA systems to answer questions about dashboard icons, which existing in-car QA systems currently do not. We created a novel dataset consisting of 408 different icons from four

different vehicle manufacturers for this specific domain and provided insights into challenges and performance in an automotive context. The impact includes improved driver safety through reduced cognitive load, as drivers can quickly access clear explanations of unfamiliar icons without manual distraction. Our work furthermore assists the development of easier to use and more powerful vehicle assistants, which benefits drivers with varying levels of automotive knowledge. Beyond driver assistance, our methodology and findings may have broader applications in evaluating LVLM performance on abstract or symbolic images across various domains, such as industrial design or medical imaging.

For future work, we plan to fine-tune LLaVA using our dataset, focusing on improving the vision encoder to better differentiate between icons and reduce hallucinations. We will explore replacing LLaVA’s original CLIP ViT-L vision encoder with more capable versions, such as those using Data Filtering Networks (DFN) (Fang et al., 2023) and quick GELU (Hendrycks and Gimpel, 2023), which have shown improved performance on ImageNet. Additionally, we aim to develop new metrics that are more responsive to hallucinations in generated image descriptions, such as visual-likeness aware named entity similarity. This approach would capture that semantically different objects (e.g., a tire cross-section and a horseshoe) may describe similar shapes, while semantically close items (e.g., a brake pedal and a brake disc) may be visually distinct. We also plan to expand the dataset by processing additional vehicle manuals and collecting more human-generated descriptions. With this study, we lay the groundwork for improved icon interpretation in conversational driver assistance systems and hope to contribute to the development of more effective and user-friendly automotive interfaces.

6 Limitations

We acknowledge several limitations with our study and current dataset. As noted, we focused entirely on vehicle manuals that were freely available and in a structured format (HTML). There are many more freely available vehicle manuals that are in PDF format; however, these are much more difficult to parse consistently, in order to extract the correct context text that goes with an image. We decided that the challenges associated with PDF

parsing were beyond current scope. Nevertheless, this broader set of manuals is a valuable data source, to which we hope to return in future. Further, since this limited the size of our dataset, we did not conduct fine-tuning of the LLaVA model. With enough additional data to preserve a respectable test set, we hope to complete an evaluation of LLaVA fine-tuning using an appropriate strategy, such as Low-Rank Adaptation (LoRA) (Hu et al., 2022).

7 Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful and encouraging comments, which have helped strengthen this paper. We also extend our gratitude to Andrew Romanof and the other volunteers for their valuable assistance with dataset creation. This work is partially supported by an NSERC Alliance grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

- Anthropic PBC. 2024. [The Claude 3 Model Family: Opus, Sonnet, Haiku](#).
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Johannes Buchner. 2024. [ImageHash: A Python Perceptual Image Hashing Module](#).
- L. Elisa Celis and Vijay Keswani. 2020. [Implicit Diversity in Image Summarization](#). *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2):139:1–139:28.
- David Chan, Austin Myers, Sudheendra Vijayanarasimhan, David Ross, and John Canny. 2023. [IC3: Image captioning by committee consensus](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8975–9003, Singapore. Association for Computational Linguistics.
- Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. [VisualGPT: Data-Efficient Adaptation of Pretrained Language Models for Image Captioning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040.
- Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. [Data Filtering Networks](#). *arXiv preprint*.
- FCA US, LLC. 2024. [Mopar Select Vehicle | Official Mopar® Site](#).
- Dan Hendrycks and Kevin Gimpel. 2023. [Gaussian Error Linear Units \(GELUs\)](#). *arXiv preprint*.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. [CLIPScore: A reference-free evaluation metric for image captioning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.
- ISO. 2021. [ISO 2575:2021](#).
- Jaguar Land Rover Limited. 2024. [Jaguar / Land Rover iGuide Online](#). [Jaguar](#) and [Land Rover](#).
- Terry K. Koo and Mae Y. Li. 2016. [A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research](#). *Journal of Chiropractic Medicine*, 15(2):155–163.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Annika Lindh, Robert J. Ross, Abhijit Mahalunkar, Giancarlo Salton, and John D. Kelleher. 2018. [Generating Diverse and Meaningful Captions](#). In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 176–187, Cham. Springer International Publishing.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual Instruction Tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.
- Mazda Canada Inc. 2024. [Owner’s Manuals for Vehicles and Connected Services | Mazda Canada](#).
- OpenAI, Josh Achiam, Stephen Adler, et al. 2024. [GPT-4 Technical Report](#). *arXiv preprint*.
- Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023. [SmallCap: Lightweight Image Captioning Prompted With Retrieval Augmentation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

3982–3992, Hong Kong, China. Association for Computational Linguistics.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. [ViperGPT: Visual Inference via Python Execution for Reasoning](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898.

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.

Volvo Canada. 2024. [Volvo Support EN-CA](#).

Yonghui Wu, Mike Schuster, et al. 2016. [Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation](#). *arXiv preprint*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). In *Eighth International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. [MiniGPT-4: Enhancing vision-language understanding with advanced large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

A Data Collection Details

We collected images of dashboard icons from 42 vehicle manuals from four different manufacturers, as shown in Table 5. All the manuals were available on the internet in HTML format ([Jaguar Land Rover Limited, 2024](#); [Volvo Canada, 2024](#); [Mazda Canada Inc., 2024](#); [FCA US, LLC, 2024](#)).

Figure 1a and 1b show screenshots of the website we developed to allow our volunteer human labelers to provide icon image descriptions.

Make	No. of Manuals	Unique Icons
Jaguar Land Rover	16	107
Volvo	15	128
Stellantis	4	130
Mazda	7	43
Total	42	408

Table 5: Summary of dashboard icons by manufacturer.

B Prompts

Three prompt types were used in the zero-shot study with multiple input levels:

- **Image and Context.** You are an AI visual assistant specialized in interpreting icons displayed on the dashboard of a vehicle. An icon communicates important information about the vehicle to the driver. For example, a particular icon may indicate that a seatbelt is not fastened. *You are seeing an image of a single dashboard icon.*

Briefly describe the dashboard icon depicted in the image, focusing on the visual content of the image and meaning of the icon. Limit your response to 2 sentences. The first sentence should describe the visual content. The second sentence should describe the icon’s meaning. Format your response as a JSON object with the following keys: ‘visual_content’, ‘meaning’. *The image has the following associated text:*

<base64-encoded icon image> <context text for icon image>

- **Image Only.** You are an AI visual assistant specialized in interpreting icons displayed on the dashboard of a vehicle. An icon communicates important information about the vehicle to the driver. For example, a particular icon may indicate that a seatbelt is not fastened. *You are seeing an image of a single dashboard icon.*

Briefly describe the dashboard icon depicted in the image, focusing on the visual content of the image and meaning of the icon. Limit your response to 2 sentences. The first sentence should describe the visual content. The second sentence should describe the icon’s meaning. Format your response as a JSON object with the following keys: ‘visual_content’, ‘meaning’.

<base64-encoded icon image>

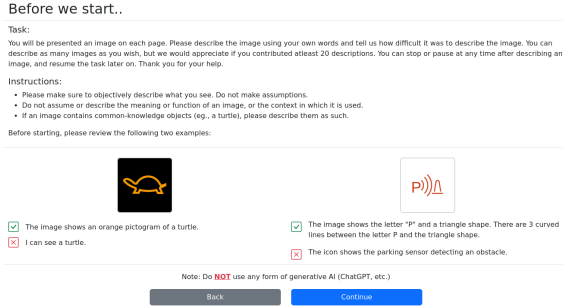
- **Context Only (Imaginary Image).** You are an AI visual assistant specialized in interpreting icons displayed on the dashboard of a vehicle. An icon communicates important information about the vehicle to the driver. For example, a particular icon may indicate that a seatbelt is not fastened. *Imagine you are seeing an image of a single dashboard icon that has an associated text description.*

Briefly describe the dashboard icon depicted in the image, focusing on the visual content of the image and meaning of the icon. Limit your response to 2 sentences. The first sentence should describe the visual content. The second sentence should describe the icon’s meaning. Format your response as a JSON object with the following keys: ‘visual_content’, ‘meaning’. *The image has the following associated text:*

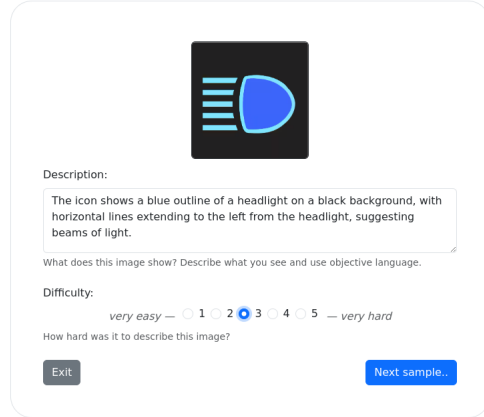
<context text for icon image>

Only one prompt type (image-and-context) was used the k -shot evaluation, with k images and ground truth descriptions appended as additional messages before the query image:

- **Image and Context.** You are an AI visual assistant specialized in interpreting icons displayed on the dashboard of a vehicle. An icon communicates important information about the vehicle to the driver. You are seeing an image of a single dashboard icon. Briefly describe the dashboard icon depicted in the image, focusing on the visual content of the image and meaning of the icon. Limit your response to 2 sentences. The first sentence should describe the visual content. The second sentence should describe the icon’s meaning. Format



(a) Instructions to volunteer labellers.



(b) Icon labelling example

Figure 1: Screenshots of website used to gather image descriptions from volunteer labellers.

your response as a JSON object with the following keys: 'visual_content', 'meaning'.

Briefly describe the dashboard icon depicted in this image. The image has the following associated text:

<base64-encoded icon image from training set> <context text for icon image> <simulated JSON response based on ground truth descriptions in training set>

...

Briefly describe the dashboard icon depicted in this image. The image has the following associated text:

<base64-encoded query image from test set> <context text for icon image>

C Statistical Analysis

C.1 Effect of Model on SBERT-Score

Since the data was not normally distributed, a Friedman test was used to analyze the effect of the model variable on the SBERT-Score. The results

Source	W	ddof1	ddof2	F	p
model	0.744	1.995	812.005	1182.612	<0.001

Table 6: Friedman test: $\text{sbert_cosine} \sim \text{model}$.

revealed a statistically significant effect between different levels of the model variable (see Table 6). Wilcoxon signed-rank tests were conducted to com-

A	B	W-val	p-corr	hedges
claude-3-5	gpt-4	12206.0	<0.001	-0.442
claude-3-5	llava	974.0	<0.001	1.231
gpt-4	llava	23.0	<0.001	1.685

Table 7: Wilcoxon Signed-Rank post-hoc tests with Holm-Bonferroni correction: $\text{sbert_cosine} \sim \text{model}$.

pare the SBERT-Scores between each pair of models and Holm-Bonferroni correction was applied

to adjust for multiple comparisons. All comparisons were statistically significant with p-values less than 0.001 (see Table 7). Specifically, the difference between Claude 3.5 and GPT-4o was significant ($W = 12206.0$, $p < 0.001$, Hedges' $g = -0.442$), indicating a moderate effect size. Both, Claude 3.5 ($W = 974.0$, $p < 0.001$, Hedges' $g = 1.231$) and GPT-4o ($W = 23.0$, $p < 0.001$, Hedges' $g = 1.685$) significantly outperformed LLaVA, with the Hedges' g -value indicating a large effect size for both comparisons.

C.2 Effect of k-shot Level on SBERT-Score

model	k-shot	mean	std	min	max	
claude-3-5	0	0.770	0.123	0.383	0.980	
	1	0.772	0.115	0.384	0.985	
	3	0.771	0.116	0.388	0.982	
	5	0.780	0.121	0.396	1.000	
	gpt-4	0	0.777	0.117	0.407	0.994
gpt-4	1	0.801	0.115	0.447	1.000	
	3	0.803	0.113	0.411	1.000	
	5	0.806	0.115	0.435	1.000	
	llava	0	0.694	0.135	0.284	0.986
		1	0.705	0.142	0.302	1.000
3		0.697	0.140	0.289	0.989	
5		0.700	0.141	0.291	1.000	

Table 8: Overview of SBERT scores by model and k-shot level.

The grand means, standard deviation, and minimum and maximum values for each model and k -level can be seen in Table 8. As data was not normally distributed, a Friedman test was conducted (see Table 9), which revealed that the difference in k -level was only statistically significant for GPT-4o ($F_{3,1159} = 58.863$, $p < 0.001$)

Model	W	ddof1	ddof2	F	p
gpt-4	0.132	2.995	1159.005	58.863	<0.001
claude-3-5	0.017	2.995	1159.005	6.596	<0.000
llava	0.003	2.995	1153.005	1.166	0.321

Table 9: Friedman test: $\text{sbert_cosine} \sim \text{model} * k\text{-shot}$

and Claude 3.5 ($F_{3,1159} = 6.596$, $p < 0.001$), but not for LLaVA ($F_{3,1153} = 1.166$, $p > 0.05$). Again, Wilcoxon signed-rank pairwise tests with

Model	A	B	W-val	p-corr	hedges
gpt-4	0	1	15356.0	<0.001	-0.388
	0	3	15048.0	<0.001	-0.432
	0	5	12980.0	<0.001	-0.466
	1	3	32616.0	0.041	-0.039
	1	5	31112.0	0.008	-0.077
	3	5	34397.0	0.131	-0.039
claude-3-5	0	1	32851.0	0.121	-0.042
	0	3	34515.0	0.291	-0.037
	0	5	26739.0	<0.001	-0.168
	1	3	37219.0	0.885	0.006
	1	5	30450.0	0.005	-0.129
	3	5	31960.0	0.045	-0.137

Table 10: Wilcoxon Signed-Rank post-hoc tests with Holm–Bonferroni correction: $\text{sbert_cosine} \sim \text{model} * k\text{-shot}$.

Bonferroni-Holm correction were used to for comparing k -shot prompting levels (A and B) for GPT-4o and Claude 3.5 (see Table 10). For GPT-4o, significant differences were observed between 0-shot and 1-shot ($W = 15356.0$, $p < 0.001$, Hedges’ $g = -0.388$), 0-shot and 3-shot ($W = 15048.0$, $p < 0.001$, Hedges’ $g = -0.432$), and 0-shot and 5-shot ($W = 12980.0$, $p < 0.001$, Hedges’ $g = -0.466$). Effects between 1-shot and 3-shot ($W = 32616.0$, $p = 0.041$, Hedges’ $g = -0.039$), and 1-shot and 5-shot ($W = 31112.0$, $p = 0.008$, Hedges’ $g = -0.077$) were also significant, but much smaller. For Claude 3.5, significant differences were observed between 0-shot and 5-shot ($W = 26739.0$, $p < 0.001$, Hedges’ $g = -0.168$). Other differences between 1-shot and 5-shot ($W = 30450.0$, $p = 0.005$, Hedges’ $g = -0.129$), and 3-shot and 5-shot ($W = 31960.0$, $p = 0.045$, Hedges’ $g = -0.137$) could also be observed, but are insignificant given the problem and the lack of a significant difference to the $k = 0$ level. Both models show varying degrees of performance improvement with increasing k -shot levels, while the biggest improvement can be consistently seen at

$k = 5$ level.

C.3 Effect of description type on SBERT-Score

description	model	mean	std	min	max
functional	claude-3-5	0.811	0.113	0.344	0.980
	gpt-4	0.829	0.108	0.373	0.994
	llava	0.718	0.131	0.182	0.986
visual	claude-3-5	0.651	0.111	0.269	0.897
	gpt-4	0.664	0.113	0.235	0.918
	llava	0.574	0.122	0.137	0.894

Table 11: Overview of SBERT cosine similarity scores by description type and model.

The grand means, standard deviation, and minimum and maximum values for each model and respective description type can be found in Table 11.

C.4 Effect of the input-level on SBERT-Score, for $k = 0$

description	input	model	mean	std	min	max
functional	context-only	claude-3-5	0.82	0.10	0.46	0.97
		gpt-4	0.83	0.11	0.40	0.99
		llava	0.72	0.14	0.18	0.98
	image-and-context	claude-3-5	0.86	0.09	0.39	0.98
		gpt-4	0.85	0.09	0.41	0.99
		llava	0.77	0.11	0.43	0.99
	image-only	claude-3-5	0.75	0.12	0.34	0.98
		gpt-4	0.80	0.11	0.37	0.99
		llava	0.67	0.11	0.23	0.95
visual	context-only	claude-3-5	0.61	0.11	0.29	0.86
		gpt-4	0.59	0.12	0.23	0.88
		llava	0.50	0.12	0.14	0.86
	image-and-context	claude-3-5	0.68	0.09	0.38	0.90
		gpt-4	0.70	0.09	0.44	0.92
		llava	0.61	0.11	0.28	0.89
	image-only	claude-3-5	0.66	0.11	0.27	0.88
		gpt-4	0.70	0.09	0.30	0.90
		llava	0.60	0.11	0.20	0.89

Table 12: Overview of SBERT-Scores by input, description type and model, for $k = 0$.

Table 12 compares the SBERT-Scores of the three models across different three input types (see Section 3.2) for both functional and visual descriptions. Note that $k = 0$ for all of these results, as few-shot prompting was not within the scope for this part of the experiment. Generally, models perform better on functional descriptions compared to visual ones. The image-and-context input consistently yields the highest mean scores across all models and description types. GPT-4o and Claude 3.5 demonstrate similar performance, often outperforming LLaVA, particularly in functional tasks. All models show improved performance when given both image and context compared to either context or image alone. The data shows considerable variability in scores, with standard devia-

tions ranging from 0.09 to 0.14 and wide ranges between minimum and maximum values, likely stemming from sample-dependent fluctuations. The

Model	W	ddof1	ddof2	F	p-unc
0 gpt-4	0.231	1.995	812.005	122.192	<0.001
1 llava	0.235	1.995	812.005	125.047	<0.001
2 claude-3-5	0.298	1.995	812.005	173.177	<0.001

Table 13: Friedman test: $\text{sbert_cosine} \sim \text{input} * \text{description_type} * \text{model}$

Friedman test results in Table 13 show significant differences across input types for all three models, with p-values < 0.001. Post-hoc Wilcoxon

Model	Description	A	B	W-val	p-corr	hedges
gpt-4	visual	c	i+c	4600.0	<0.001	-1.067
	visual	c	i	6276.0	<0.001	-1.015
	visual	i+c	i	37276.0	0.122	0.047
	functional	c	i+c	31423.0	0.002	-0.191
	functional	c	i	30245.0	<0.001	0.259
	functional	i+c	i	23192.0	<0.001	0.474
llava	visual	c	i+c	9540.0	<0.001	-0.974
	visual	c	i	14501.0	<0.001	-0.860
	visual	i+c	i	36691.0	0.035	0.093
	functional	c	i+c	22527.0	<0.001	-0.461
	functional	c	i	28700.0	<0.001	0.384
	functional	i+c	i	13745.0	<0.001	0.970
claude-3-5	visual	c	i+c	11767.0	<0.001	-0.737
	visual	c	i	23117.0	<0.001	-0.466
	visual	i+c	i	24519.0	<0.001	0.238
	functional	c	i+c	23716.0	<0.001	-0.321
	functional	c	i	19376.0	<0.001	0.643
	functional	i+c	i	8579.0	<0.001	0.974

Table 14: Wilcoxon Signed-Rank post-hoc tests with Holm-Bonferroni correction: $\text{sbert_cosine} \sim \text{input} * \text{description_type} * \text{model}$.

signed-rank tests with Holm-Bonferroni correction, as shown in Table 14 confirmed these performance variations across input types and description tasks. For visual descriptions, all models show significant improvements when using image-and-context or image-only inputs compared to context-only, with generally larger effect sizes for image-and-context. In functional descriptions, image-and-context consistently outperforms other input types, while the relationship between context-only and image-only inputs varies by model. Claude 3.5 demonstrates the most consistent pattern across both description types, with significant differences and substantial effect sizes between all input type pairs.

C.5 Effect of image difficulty on SBERT-Score

We evaluated the degree to which each model-generated description relied on the content of the image versus the manual context text. Using a zero-shot strategy, each model was prompted to return

one visual description and one functional description of each icon image. Three types of prompts were used in this evaluation: (1) the model was supplied with both the encoded icon image and context text; (2) only the encoded icon image was supplied to the model; and (3) only the context text was supplied. For this third case, the model was asked to imagine an image that matched the supplied context text and then return a visual description of the imagined image.

We found a significant linear trend in the effect of the difficulty on SBERT-Scores ($\beta = -0.133$, $p < 0.001$), indicating a general decrease in performance as difficulty increases. This weak monotonic relationship was confirmed using Spearman’s rank correlation ($\rho = -0.254$, $p < 0.001$).

As image complexity increases, model performance decreases, with a weak to moderate negative correlation between image difficulty and description accuracy.

C.6 Human Evaluation

C.6.1 Design Details

The human evaluation was designed using a balanced incomplete block approach to assess 60 samples, representing 15% of the dataset. Six participants were recruited for the study, with each participant evaluating a subset of the samples. The samples were divided into four blocks, each containing 15 samples. As shown in Figure 2, the evaluation

		Samples				
		~	01..15	16..30	31..45	46..60
Participant	A					
	B					
	C					
	D					
	E					
	F					

Figure 2: Balanced Incomplete Block Design for Human Evaluation

process followed a pattern where each participant rated a specific set of samples across two different blocks. The block design balanced the distribution of samples among participants, with each participant assessing 30 samples in total, resulting in three independent ratings per sample. The overlapping blocks were chosen to mitigate potential biases that could arise from assigning all samples to a single rater. This proved effective, as discussed in the following subsection.

C.6.2 Results

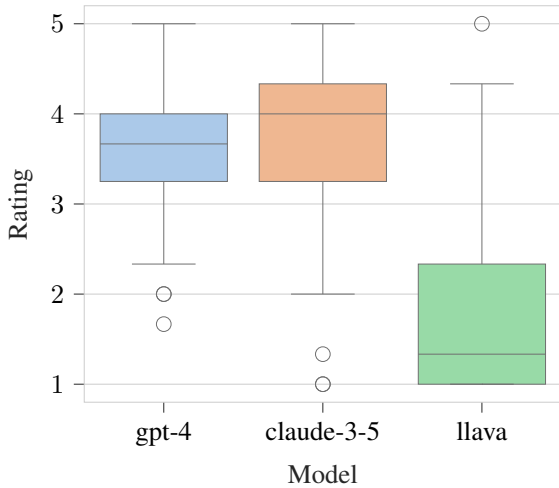


Figure 3: Average Human Evaluation Ratings by Model

A linear mixed-effects model analysis was conducted to investigate the effect of different models on the human evaluation ratings, the distribution of which is shown in Figure 3. Model and participant were treated as fixed effects, while the four blocks were modeled as random effects. Additionally, individual images were accounted for using varying coefficients. The regression results in Table 15 reveal significant differences in ratings across models and participants. Neither Claude

Mixed Linear Model Regression Results						
Model:	MixedLM	Dependent Variable:	rating			
No. Observations:	720	Method:	REML			
No. Groups:	4	Scale:	0.9765			
Min. group size:	180	Log-Likelihood:	-1058.0983			
Max. group size:	180	Converged:	Yes			
Mean group size:	180.0					
	Coef.	Std.Err.	z	P> z	[0.025	0.975]
Intercept	3.874	0.153	25.375	0.000	3.575	4.174
C(model)[T.claude-3-5]	0.189	0.104	1.813	0.070	-0.015	0.393
C(model)[T.gpt-4]	0.106	0.104	1.013	0.311	-0.099	0.310
C(model)[T.llava]	-1.572	0.104	-15.094	0.000	-1.776	-1.368
C(participant)[T.2]	0.040	0.141	0.287	0.774	-0.235	0.316
C(participant)[T.3]	-0.593	0.152	-3.914	0.000	-0.890	-0.296
C(participant)[T.4]	-0.457	0.142	-3.218	0.001	-0.735	-0.179
C(participant)[T.5]	-0.542	0.141	-3.855	0.000	-0.817	-0.266
C(participant)[T.6]	-0.928	0.142	-6.525	0.000	-1.207	-0.649
block Var	0.025	0.038				
sample Var	0.189	0.054				

Table 15: Linear Mixed Effects Model: rating ~ model + participant.

3.5 nor GPT-4o show statistically significant differences ($p > 0.05$). However, LLaVA demonstrates a highly significant negative effect ($p < 0.001$), with substantially lower ratings compared. Participant effects are evident, with all participants except participant 2 rating significantly lower than the reference participant. Notably, the balanced incomplete block design with four overlapping blocks

(see Section 4.2) proved effective in mitigating the impact of varying participant rating habits, as indicated by minimal variability between experimental blocks. While substantial variability was observed between individual images, this was anticipated given the varying degrees of difficulty in the image set.

C.6.3 Inter-Rater Agreement

The inter-rater agreement for our human evaluation was assessed using intraclass correlation coefficients (ICC) and visual inspection of the ratings. Figure 4 presents the average scores for each participant and model, revealing consistent trends across raters. The ICC analysis shows strong agree-

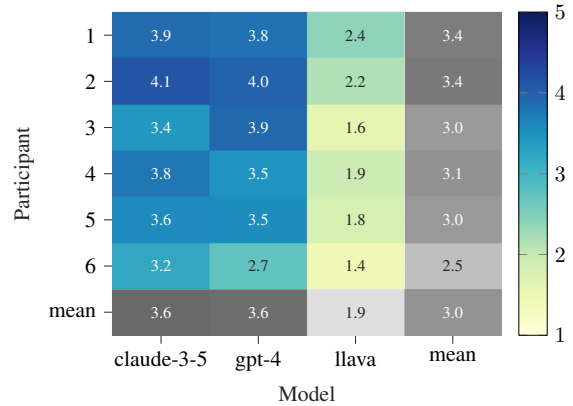


Figure 4: Ratings by Participant and Model

ment among raters. The ICC(3,k) value of 0.987 (95% CI: [0.95, 1.0]) indicates excellent agreement for average fixed raters, and the ICC(1) value of 0.771 (95% CI: [0.43, 0.98]) suggests good agreement even for single raters. These high ICC values demonstrate reliable consensus among participants in their assessments. Examining the ratings, we observe that all participants consistently ranked the models in the same order: Claude 3.5 slightly outperforming GPT-4, with LLaVA receiving notably lower scores. While there are some variations in individual scoring patterns (e.g., participant 6 generally gave lower scores), the overall trends remain consistent.

C.6.4 Correlation with Automated Metrics

To evaluate the performance of the different automated metrics in predicting human judgments of image description quality, we conducted a Spearman’s rank correlation analysis with bootstrapping. This approach allowed us to assess the robustness of the correlation between each metric and human ratings while mitigating the impact of individual

samples. We performed 1000 bootstrap iterations, calculating the Spearman’s rho for each metric in each iteration. The outcome of this analysis can

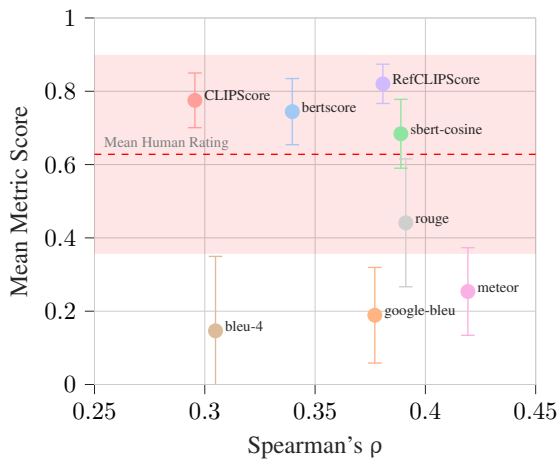


Figure 5: Correlation of Automated Metrics with Human Ratings

be seen in Figure 5, all correlations were of positive, weak to moderate degree. The results revealed that Meteor and ROUGE demonstrated the highest correlations with human ratings ($\rho = 0.439$ and $\rho = 0.417$ respectively), followed closely by the SBERT-Score ($\rho = 0.395$). However, when considering both correlation strength and the interpretability of the metric scores, SBERT-Score emerged as the preferred choice. Unlike Meteor and ROUGE, which produced lower average scores with high standard deviations (0.26 ± 0.12 and 0.44 ± 0.17 respectively), SBERT-Score yielded a mean score of 0.68, which aligned much more closely with the average percentage of the human evaluation at 0.63 (calculated as $\frac{3.14}{5} = 0.63$). This balance between a comparably high correlation with human judgements and producing easily interpretable similarity scores made the SBERT-Score most suitable for evaluating image descriptions in our context, where both alignment with human perception and general similarity levels were of importance.

D Additional Results


Icon	Context
	See DIESEL EXHAUST FLUID (DEF) (RED)
Ground Truth Visual Description	
This red dashboard icon depicts two horizontal lines representing emissions from the vehicle. Two drops of fluid are shown on top of the horizontal lines with a wavy line at the bottom.	
k-shot	Visual Description Generated by LLaVA
1	The icon is a red warning sign with the words 'DIESEL EXHAUST FLUID' and an image of a car or vehicle
3	The dashboard icon displays a red fuel canister symbol with wavy lines above it, signifying emission
5	The red dashboard icon shows an abstract representation of a vehicle with two red stars above it, suggesting energy or combustion.

Table 16: Example of LLaVA generations with increasing k for k-shot prompting.