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

James Fletcher[1], Nicholas Dehnen[1], Seyed Nima Tayarani Bathaie[1], Aijun An[1], Heidar Davoudi[2], Ron di Carlantonio[3], Gary Farmaner[3]

[1]Lassonde School of Engineering, York University, Toronto, Canada; [2]Faculty of Science, Ontario Tech University, Oshawa, Canada; [3]iNAGO Co., Toronto, Canada

Introduction

Vehicle dashboard icons convey critical information to drivers, who must quickly understand these symbols to take appropriate action. But many drivers are unfamiliar with these icons.

iNAGO's **netpeople** is a voice-based virtual assistant for automotive drivers. netpeople's text-based knowledge base (KB) currently lacks icon descriptions and it struggles with icon-related inquiries.

Objective: Automatically generate text descriptions for icon images, enabling netpeople's KB to include questions and answers about dashboard icons.

Challenges:

1. Existing image description systems train on natural images, whereas icons are drawings.
2. Understanding an icon's function, beyond its visual description, requires context from the vehicle manual.
3. No suitable labeled dataset currently exists.
4. Many different metrics are used to evaluate generated descriptions. Need to identify the best ones for this use case.

Ground Truth Functional Description:
The icon indicates that the vehicle's driver condition monitor system has detected that the driver is presenting signs of high fatigue levels.

Ground Truth Visual Description:

1. An amber coffee mug on a coaster. Wavy vertical lines indicate steam rising from the coffee mug.
2. An orange-coloured cup placed on a saucer. The steam is coming out of the cup.
3. A cup and saucer. Three wavy lines above the cup show that the cup contains a hot drink.

Figure 1. An example icon. Context: "See Driver Condition Monitor (Amber)"

Dataset Development

Data: We processed 42 HTML vehicle manuals (available online) and collected 408 unique dashboard icon images. We also extracted context text from the manuals associated with each icon image.

Ground truth: For each icon, we used human volunteers to generate two types of descriptions:

1. Up to three diverse **visual descriptions** of recognizable image components. We created a web interface to collect these descriptions from 28 volunteer annotators.
2. A single **functional description** based on manual text.

The visual and functional descriptions form the question and answer, respectively, in netpeople's KB.

Icon	Model	Generated Visual Description	SBERT Score	Human Eval.
	GPT-4o	This amber dashboard icon depicts a cup of steaming hot beverage, such as coffee or tea.	0.66	4
	Claude 3.5	This amber dashboard icon depicts a coffee cup with steam rising from it.	0.67	3.7
	LLaVA	The icon depicts a stylized representation of a cup with steam rising from it.	0.69	3.7
	GPT-4o	This dashboard icon depicts a vehicle headlight with five horizontal lines extending to the left, indicating the light beams.	0.70	3.7
	Claude 3.5	This dashboard icon depicts a vehicle headlight with five horizontal lines extending to the left, indicating the light beams.	0.65	4.7
	LLaVA	The icon depicts a headlight *with a snowflake inside, representing icy road conditions* while the high beam is on.	0.45	1.0

Table 1. Examples of visual descriptions generated by the models (using 3-shot prompting), alongside their SBERT cosine similarity and human evaluation scores.

Generating Icon Descriptions

Models: Three state-of-the-art pre-trained Large Vision-Language Models (LVLMs) were used to generate visual and functional descriptions for each icon image: **GPT-4o**, **LLaVA-NEXT:34b**, **Claude 3.5 Sonnet**.

Prompts: Each model was provided with an icon image plus its context text and an appropriate prompt. Generated visual and functional descriptions were collected separately.

We tried both few-shot and zero-shot prompts. For few-shot, we selected k examples from a 20-icon training set that were closest to the query icon by comparing image hashes (Hamming distance).

Evaluation: We used several types of automatic metrics to evaluate the model-generated descriptions against ground truth. We also randomly selected 60 test icons and asked six people to rate the generated visual descriptions on a one to five Likert scale.

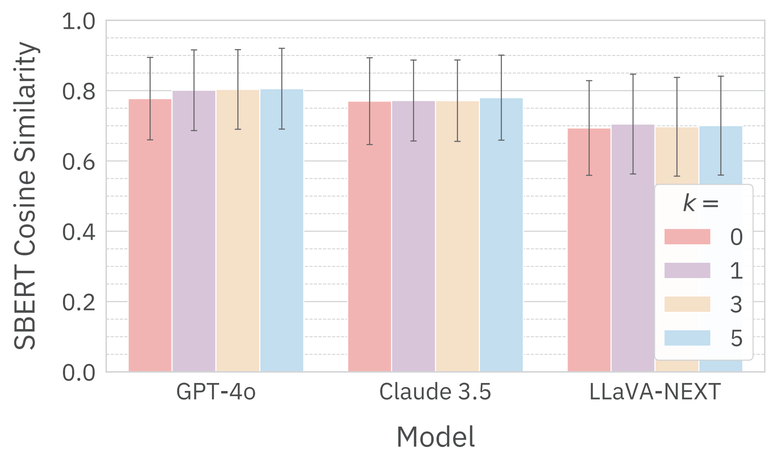

Figure 2. Bar plot of mean SBERT Cosine Similarity scores by model and k-level.

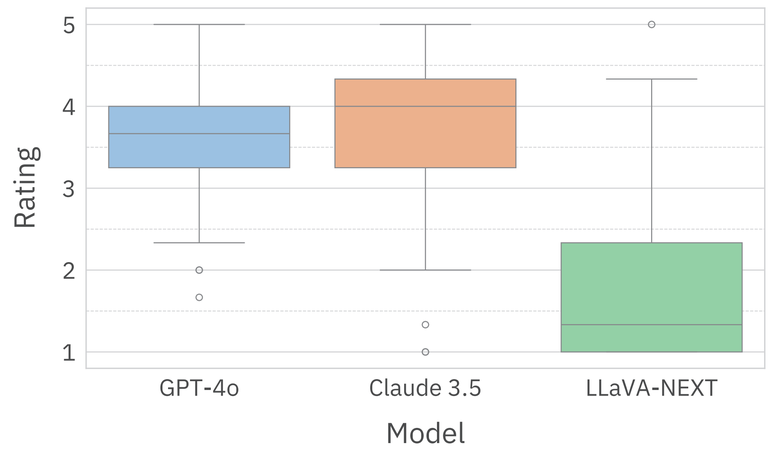

Figure 3. Box plot of human evaluation results by model.

Results and Discussion

Overall: Most metrics produced the same ranking of the models for both description types:

- GPT-4o performed best
- Followed closely by Claude 3.5 Sonnet
- LLaVA performed relatively poorly

Metrics and Performance: We found SBERT cosine similarity to have the highest correlation with human ratings. We used a Friedman test to assess significance of the following results:

1. GPT-4o performs significantly better than the other two models and Claude 3.5 is significantly better than LLaVA.
2. Few-shot prompting significantly improves performance for GPT-4o and Claude 3.5, with 5-shots showing the largest effect. LLaVA does not benefit from few-shot prompting.
3. For all models, performance on visual descriptions is significantly worse than on functional descriptions This may be influenced by the context, each model's vision capabilities, and the variability in visual descriptions (e.g., object names vs. simple geometric shapes).
4. Human evaluators score GPT-4o and Claude 3.5 significantly better than LLaVA. Among the automatic metrics, SBERT cosine similarity is most consistent with human ratings.

Conclusion

We have presented a new application of large vision-language models for interpreting and describing vehicle dashboard icons. Our contributions include:

- The novel task of automatic generation of visual and functional descriptions of automotive icons.
- A novel dataset consisting of 408 different icons from four different vehicle manufacturers for this specific domain.
- Insights into challenges and performance in an automotive context.
- Both automatic and human evaluation revealed strong performance from GPT-4o and Claude 3.5, yet all models suffer from hallucinations for less common symbols.

The impact of our work includes improved driver safety through reduced cognitive load and assists the development of easier to use and more powerful vehicle assistants. Beyond driver assistance, our methodology and findings may have broader applications in evaluating LVLM performance on abstract or symbolic images across various domains.

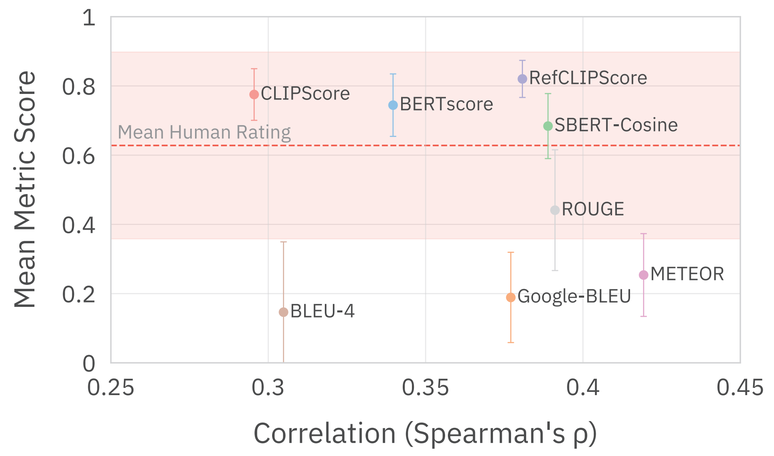

Figure 4. Scatter plot showing mean and ± 1 standard deviation for each automatic metric (y-axis) and with results from human evaluation (x-axis)

1. "Cross-section of a tire.."
2. "U-shape with a flat bottom.."
3. "Lyre without strings.."

Figure 5. Excerpts of visual descriptions from three different human annotators, each describing the outer shape of the symbol in their own way.

Impact & Future Directions

Implementation Insights:

- Multi-modal input (image + context) consistently outperforms single-modal approaches
- Few-shot prompting improves performance for GPT-4o and Claude 3.5
- SBERT cosine similarity shows strongest correlation with human judgment

Real-World Impact:

- Enables real-time icon interpretation for drivers
- Reduces cognitive load during vehicle operation
- Improves accessibility for drivers unfamiliar with dashboard symbols
- Supports development of more intuitive vehicle interfaces

Research Roadmap:

- Expand dataset with PDF processing
- Fine-tune vision encoders for improved icon recognition
- Develop hallucination- and visual-likeness-aware evaluation metrics
- Integrate with production vehicle systems

This work bridges the gap between technical advancement and practical automotive safety applications while laying groundwork for future improvements in vehicle-driver interaction.

References

1. iNAGO. 2024. netpeople Assistant Platform.
2. Johannes Buchner and Chris Pickett. 2021. ImageHash: A Python Perceptual Image Hashing Module.
3. OpenAI, Josh Achiam, Steven Adler, et al. 2024. GPT-4 Technical Report.
4. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. Advances in Neural Information Processing Systems, 36:34892–34916.
5. Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Acknowledgements

iNAGO
NSERC CRSNG
LASSONDE SCHOOL OF ENGINEERING
YORK UNIVERSITÉ UNIVERSITY