Automated Text Descriptions for Vehicle Dashboard Icons Using Large Vision-Language Models

James Fletcher[1], Nicholas Dehnen[1], Seyed Nima Tayarani Bathaie[1], Heidar Davoudi[2], Aijun An[1]

[1]Dept. of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University; [2]Faculty of Science, Ontario Tech University

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. The image shows an amber coffee mug on a coaster. Wavy vertical lines indicate steam rising from the coffee mug.

2. The pictogram depicts an orange coloured cup placed on a saucer. The steam is coming out of the cup.

3. This amber dashboard icon depicts 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.

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**.

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.

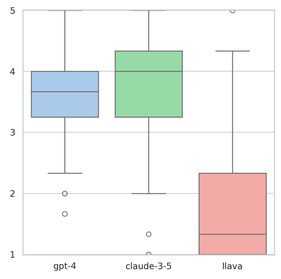


Figure 3. Box plot of human evaluation results for each model (GPT-4o, LLaVA, Claude 3.5).

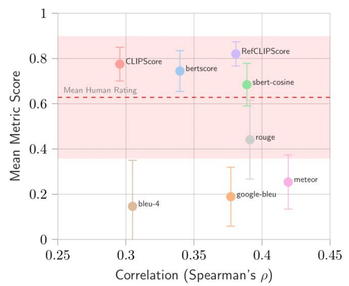


Figure 4. Scatter plot showing:

- *Vertical axis*: mean ±1 standard deviation for each automatic metric
- *Horizontal axis*: correlation of automatic metrics with results from human evaluation

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	0.70	3.7
	LLaVA	The icon depicts a stylized representation of a cup with steam rising from it	0.69	4.0

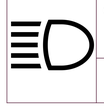
Icon	Model	Generated Visual Description	SBERT Score	Human Eval.
	GPT-4o	This dashboard icon depicts a vehicle headlight with five horizontal lines extending to the left, indicating the light beams	0.70	3.7
	LLaVA	The icon depicts a headlight *with a snowflake inside, representing icy road conditions* while the high beam is on	0.44	1.0

Table 1. Two examples of visual descriptions generated by GPT-4o and LLaVA (3-shot prompting). SBERT cosine similarity scores and human evaluation scores are also shown.

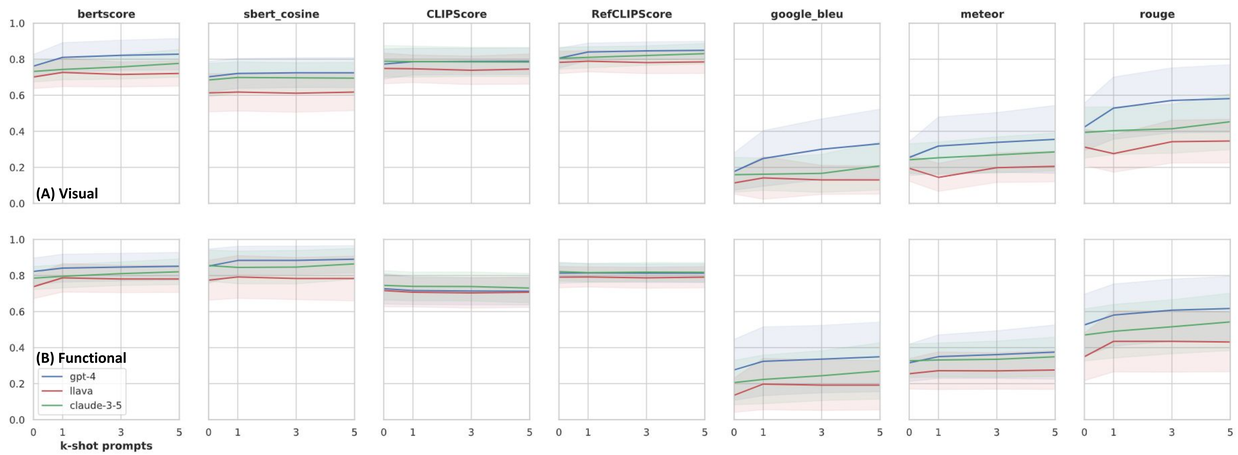


Figure 2. Comparison of generated visual (A) and functional (B) descriptions for the three models (GPT-4o, LLaVA, Claude 3.5) as compared against ground truth descriptions using seven automated metrics (BERT-score, cosine similarity with SBERT, CLIP-score, RefCLIP-score, Google-BLEU-4, METEOR, ROUGE). Results are shown for zero-shot and few-shot prompting strategies (*k*= 0, 1, 3, 5). The solid line indicates the mean value and the shaded bands indicate ±1 standard deviation about the mean.

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
- LLaVA performed relatively poorly

The exception was CLIP and RefCLIP, which ranked Claude 3.5 higher than GPT-4o on functional descriptions. This is likely because CLIP has no access to the context text when it generates the reference description.

Best Metric: We found SBERT cosine similarity to be the best automatic metric. We used a Friedman test to assess significance (since the data were not normally distributed).

1. GPT-4o performs significantly better than the other two models and Claude 3.5 is significantly better than LLaVA ($p < 0.001$).
2. Few-shot prompting significantly improves performance for GPT-4o and Claude 3.5, with 5-shots showing the largest effect ($p < 0.001$). LLaVA does not benefit from few-shot prompting ($p > 0.05$).
3. For all models, performance on visual descriptions is significantly worse than on functional descriptions ($p < 0.001$). 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 ($p < 0.001$). Among the automatic metrics, SBERT cosine similarity is most consistent with human ratings.

Conclusions

- The application of LVLMs to generation of vehicle dashboard icon descriptions is novel.
- The vehicle dashboard icon dataset we created is also new and distinguishes between two types of icon image descriptions: visual and functional.
- The generated descriptions will enable iNAGO's netpeople to answer questions about dashboard icons.
- Both automatic and human evaluation revealed strong performance from GPT-4o and Claude 3.5, while LLaVA performed poorly.
- **Future Work**: Our dataset has good coverage of vehicle manuals from four manufacturers but many more remain unrepresented. We hope to expand the dataset by processing additional PDF manuals and collecting more human-generated descriptions.

References

1. iNAGO. 2024. netpeople Assistant Platform. Available at: www.inago.com/products
2. Johannes Buchner. 2024. ImageHash: A Python Perceptual Image Hashing Module. GitHub. Available at: github.com/JohannesBuchner/imagehash
3. OpenAI. 2024. GPT-4 Technical Report. Preprint, arXiv:2303.08774. Available at: arxiv.org/abs/2303.08774
4. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. Preprint, arXiv:2304.08485. Available at: arxiv.org/abs/2304.08485
5. Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Available at: www.anthropic.com/claude-3-model-card

Acknowledgements

iNAGO

LASSONDE SCHOOL OF ENGINEERING | YORK U