Scene Understanding Maps: Predicting Most Frequently Fixated Object during Scene Description with Multi-Modal Large Language Models

Shravan Murlidaran (smurlidaran@ucsb.edu)[1], Ziqi Wen[2], Jonathan Skaza[3], and Miguel P. Eckstein[1]

[1]Dept. of Psychological and Brain Sciences, [2]Dept. of Computer Science, [3]Dept. of Dynamical Neuroscience

University of California, Santa Barbara

VIU vision ● image understanding

Introduction

- Over the last decade, image computable models such as saliency (GBVS[1]) or Deepgaze[2] have been developed to predict locations people might fixate when viewing naturalistic scenes.
- A recent technique (Scene Understanding Map (SUM)) developed by Murlidaran et al.[3] showed that people fixate on objects critical to understanding the scene. They digitally removed objects from a scene and quantified the impact of the removal by comparing the similarity of descriptions from humans with and without the object's presence.
- Here, we build upon their method by automating various components of their procedure. To this end, we use the Winograd images developed by Murlidaran et al. to compare the effect of automating various components on the agreement of the most critical object with the original method. We also compare the ability of all the methods to predict the most fixated object on a scene description task.

References

1. Harel, J., Koch, C. and Perona, P., 2006. Graph-based visual saliency. *Advances in neural information processing systems, 19*.
2. Kümmerer, M., Wallis, T.S. and Bethge, M., 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*.
3. Murlidaran, S., & Eckstein, M. P. (2024). Eye Movements during Free Viewing Maximize Scene Understanding. *Journal of Vision, 24*(10), 1189-1189.
4. Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... & Kivlichan, I. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
5. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2023). *Grounding dino: marrying dino with grounded pre-training for open-set object detection. arXiv abs/2303.05499 (2023)*.
6. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., ... & Zhang, L. (2024). Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.
7. Zhuang, Junhao, et al. "A task is worth one word: Learning with task prompts for high-quality versatile image inpainting." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

Winograd Pair Example

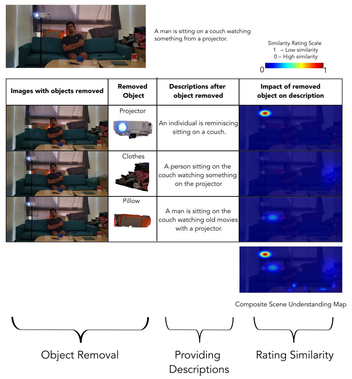
Image pairs that look visually very similar but are described differently

Description: A man is sitting on a couch watching something from a projector.

Descripton: A man is sitting on a couch while folding laundry.

Methods

SUM map generation has three stages



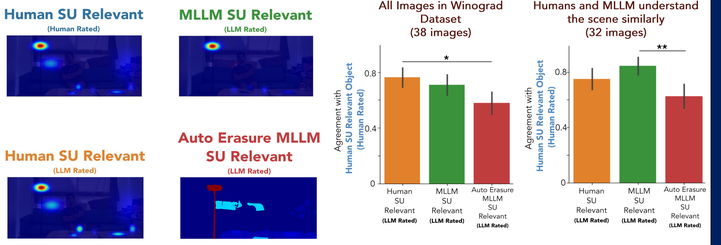
Automating the SUM generation process at different stages

Stage \ SUM Method	Original Method: Human SU Relevant (Human Rated)	Human SU Relevant (LLM Rated)	MLLM SU Relevant (LLM Rated)	Auto Erasure - MLLM SU Relevant (LLM Rated)
Object Removal	Human	Human	Human	Auto Segmentation and in-painting
Providing Descriptions	Human	Human	MLLM	MLLM
Rating Similarity	Human	LLM	LLM	LLM

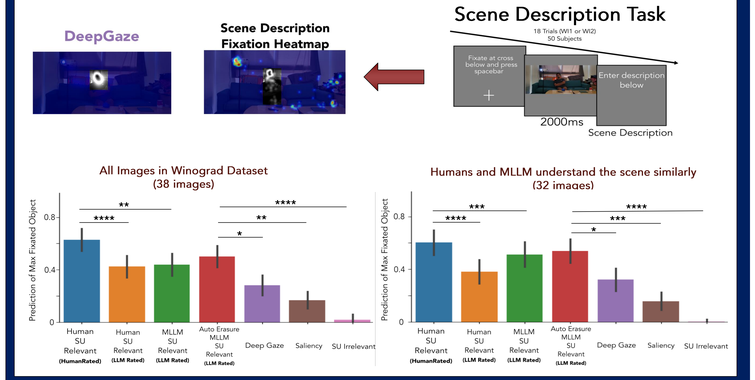
LLM - Large Language Model **MLLM - Multi-Modal LLM**

Results

Agreement with the original method to generate the SUM map



Predicting Maximum Fixated Object in Scene Description Task



Auto Erasure Procedure

Prompt GPT4o[4] to get a list of objects

"table with metal chairs, man in blue sweatshirt, black backpack, red drink container, person in black shirt and shorts, building with windows and doors

Grounded DINO[5] (detection) & Grounded SAM[6] (segmentation)

Detection

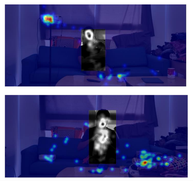
Segmentation

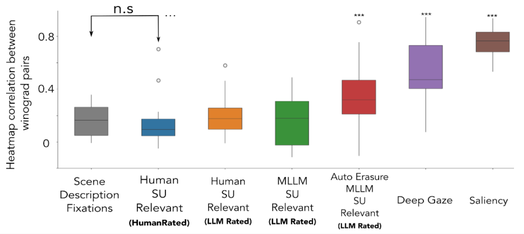
PowerPaint[7] (Image Inpainting Model (removes the segmented region)

Results

Heatmap Correlations across Winograd Pairs

Example scene description fixation heat maps for a Winograd pair





Limitations for generalization

Deleting very large objects can induce artifacts that can impact the model's understanding of the scene, though it may not be a critical object

The person appears to be taking a break after exercising with dumbbells in a living room setting.

Couch Removed

A person is examining a small diorama on the floor in a room that also contains a keyboard and guitar.

The model's understanding of the scene may not be similar to how humans understand the scene

Human Description: A man is bending to look for a tennis ball under sofa.

MLLM Description: A person is bending over, possibly looking at or interacting with something on a table in a living room.

4. Conclusion

- Scene Understanding Maps perform well at predicting the most fixated object during a scene description task.
- MLLMs can be used to automate the generation of Scene Understanding Maps. Although they have certain limitations, their performance is still at par with the original SUM maps