# Foveated Multi-Modal Large Language Model Maps to Predict Time to Understand Scenes

Ziqi Wen, Jonathan Skaza, Shravan Murlidaran, William Wang, Miguel Eckstein

University of California, Santa Barbara

VIU vision image understanding

## Motivation & Introduction

- The rapid improvement of Vision Language Models (VLM) makes the idea of image-computable model for predicting Response Time (RT) for understanding a scene possible.
  - VLM can generate scene descriptions for any image input
  - Similarity metrics based on language embeddings allow quantitative evaluations of the similarity of a human description of a scene to a gold standard description.
- What cause the main bottleneck in human scene understanding?
  - **Hypothesis**: The interaction between the foveated nature of the human visual system and the spatial distribution across the image of the visual information is critical to comprehending the scene.

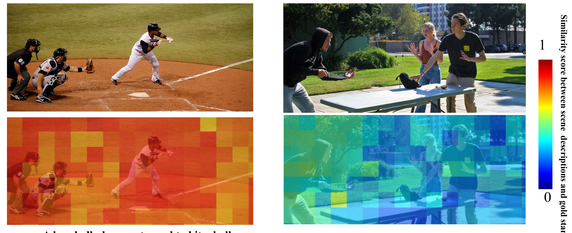








Figure 1: Comparison between low-effort and high-effort scene understanding.

## Psychophysics

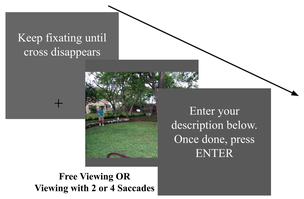


Figure 2: Overview of human psychophysics

- (1) Response Time Study:
  Participants ($N = 17$) will do the free viewing until they understand the scene, and then type in the scene descriptions. Response time and collected eye-tracking data, including the total number of saccades, are collected.
- (2) Saccade-Limited Presentation Study:
  Each scene was displayed for a restricted number of eye movements, either 2 or 4 saccades. Participants ($N = 16$) were then instructed to provide descriptions of the scene based on what they observed.

## Method: Foveated Scene Understanding Map (F-SUM)

- **Possible Baseline**:
  - GPT4o prompted Q&A: Prompting GPT4o to directly assess the difficulty for understanding a scene
  - Clutter Metrics: Feature Congestion [6] & Subband Entropy [7]
  - Image Complexity
    - Image Complexity (ICNet): DNN-based methods to predict human-annotated visual complexity score. [1]
    - Meaningful Image Complexity: Computational methods that calculate image complexity based on algorithmic information theory. [2]
  - Language Uncertainty: Measuring the model's predictive entropy over possible language responses. [3]
- **Proposed Methods**:

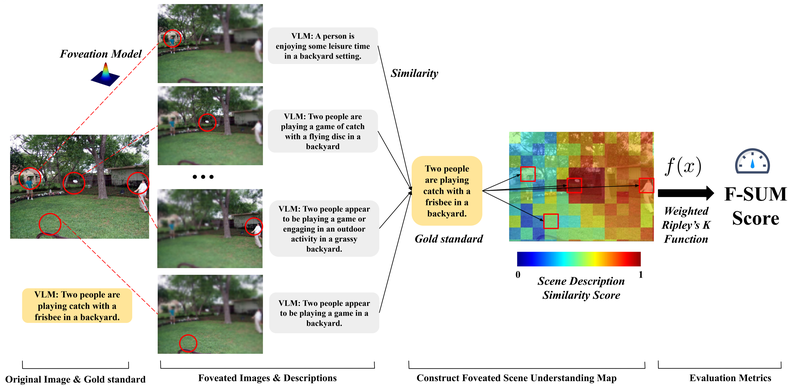


Figure 3: Overview of our Foveated Scene Understanding Map

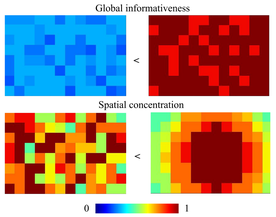


Figure 4: Criteria of Weight Ripley's K Function

- Foveated Scene Descriptions: We sample 108–136 fixation points. For each point, a foveation model is applied to simulates the fall-off in human visual acuity with eccentricity [4]. A VLM then generates a description for each foveated image.
- The F-SUM is defined as a 2D matrix $M$, where each entry $M_{m,n}$ corresponds to a grid location $(m, n)$ and is computed as:

$$M_{m,n} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{g_i \cdot f_{m,n,j}}{|g_i|\,|f_{m,n,j}|}$$

$g_i$ denotes the embedding of the $i$-th gold standard description, and $f_{m,n,j}$ denotes the embedding of the $j$-th foveated description at $(m, n)$.

- Evaluation Metrics:The design of Weighted Ripley's K (inspired by [5]) Function is driven by the two criteria:
  - Global informativeness: Scenes with low (high) understanding values at all possible fixation points should receive a low (high) understanding score.
  - Spatial concentration: Scenes in which semantic understanding is spatially clustered should score higher than scenes that have such information spatially dispersed.

$$K(r) = \frac{1}{N} \sum_{p,q \in \mathcal{P}, p \neq q} w_p w_q \cdot \mathbf{1}(r - 1 \leq d_{pq} \leq r)$$

$$S = \frac{\sum_{r=1}^{R} K(r) \cdot w_r}{\sum_{r=1}^{R} w_r}$$

$\mathcal{P}$ is the set of coordinates of elements in the matrix, $p$ and $q$ are two pixels in $\mathcal{P}$, with coordinates $(i_p, j_p)$ and $(i_q, j_q)$. $d_{pq}$ is the Euclidean distance between $p$ and $q$. $w_p$ and $w_q$ are the values of pixels $p$ and $q$. $w_r$ is the weight for distance $r$. $R$ is the maximum distance ($R = 10$ in our experiments). $S$ will be further normalized to range from 0 to 1 and let a higher value means the scene is more difficult to understand.

## Results

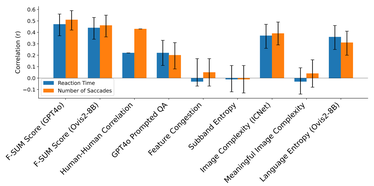


Figure 5: Correlation of each metric with the RT and the number of saccades

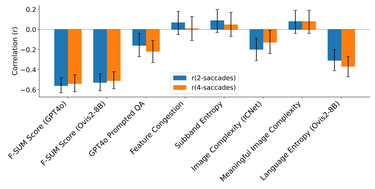


Figure 6: Correlation of each metric with the similarity of human response to gold standard under limited saccades

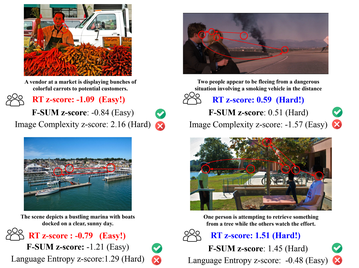


Figure 7: The difference between what F-SUM could capture and what Image Complexity and Language Entropy could capture

- As shown in Fig.5: The F-SUM Score significantly outperforms all baseline metrics in predicting both response time and saccade count ($p < .005$).
- As shown in Fig.6: The F-SUM Score significantly ($p < .0001$) outperforms all baseline metrics in predicting participants' ability to extract and articulate the gist of a scene when access to visual information was limited by spatial viewing constraints.
- Image complexity can be biased by the sheer number of visual elements—such as the densely packed carrots (top-right example in Fig.7) in the upper-left image—which may not reflect true cognitive demands.
- Language entropy captures the diversity of sampled descriptions but may miss the underlying effort required to generate them. In the bottom-left example in Fig.7, captions consistently mention a person reaching into a tree while others look on. Although the descriptions seem coherent and varied, they require multiple eye movements to integrate dispersed visual information.

## Acknowledgements

TODO

## Reference

[1] Tinglei Feng, Yingjie Zhai, Jufeng Yang, Jie Liang, Deng-Ping Fan, Jing Zhang, Ling Shao, and Dacheng Tao. Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8577–8593, 2022.
[2] Louis Mahon and Thomas Lukasiewicz. Minimum description length clustering to measure meaningful image complexity. *Pattern Recognition*, 145:109889, 2024.
[3] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.
[4] Jeffrey S Perry and Wilson S Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *Human vision and electronic imaging VII*, volume 4662, pages 57–69. SPIE, 2002.
[5] Brian D Ripley. The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266, 1976.
[6] Ruth Rosenholtz, Yuanzhen Li, Jonathan Mansfield, and Zhenlan Jin. Feature congestion: a measure of display clutter. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 761–770, 2005.
[7] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. Measuring visual clutter. *Journal of vision*, 7(2):17–17, 2007.