Contextual Semantic Relevance Predicting Human Visual Attention

Kun Sun[1,2] and Rong Wang[2,3]

[1] Tongji University, China | [2] University of Tübingen, Germany | [3] University of Stuttgart, Germany

Emails: kunsun@tongji.edu.cn; rong.wang@uni-tuebingen.de

Abstract: Understanding how humans process visual information is crucial for advancing cognitive science and developing intelligent systems. Semantic relevance, which refers to the inherent meaning of objects and their contextual relationships within a scene, plays a crucial role in predicting human visual attention. Traditional saliency models often focus mainly on low-level perceptual features or treat visual and linguistic information as separate modalities (Hayes & Henderson, 2021; Hwang et al., 2011), overlooking the potential synergy between vision and language in guiding attention. This study introduces new metrics of contextual semantic relevance that combine vision-based and language-based perspectives to examine their influence on human visual processing.

Using a large-scale eye-tracking dataset from the "Human Attention in Image Captioning" corpus (He et al., 2019), which includes fixation duration and count data for 1,000 diverse images, we developed metrics that capture how target objects relate to their visual and semantic context. We employed two state-of-the-art vision models, DINOV2 (facebook/dinov2-base) trained purely on visual features without linguistic supervision, and CLIP (openai/clip-vit-large-patch14) which aligns visual and linguistic spaces, to compute vision-based metrics. Our vision-based metrics quantify the visual similarity between a target object and surrounding objects (objs_vissim), between the object and the overall scene (obj_image_vissim), and across all scene elements (overall_vissim). Our language-based metrics, computed using Sentence Transformer embeddings (Reimers & Gurevych, 2019), measure semantic similarity between object labels and image captions at multiple levels: individual words (words_semsim), full sentences (sent_semsim), and comprehensive semantic relationships (overall_semsim). Fixation measures were recorded for specific labeled objects within each image (e.g., “bottle”, “baby”, “lemon”), capturing both fixation duration and count for each target object. The computation of these metrics is illustrated in Fig. 1. Unlike prior approaches that examined these modalities separately (Sun & Liu, 2025), we developed a combined metric (total_vissem_sim) that integrates visual and linguistic information through linear combination with equal weighting ($\alpha = \beta = 1$), a principled baseline that empirical testing showed performs comparably to optimized weights. These metrics serve dual purposes: as computational proxies that predict human attention patterns, and as tools to investigate how vision and language systems process semantic information differently, revealing insights about multimodal cognitive processing.

We employed Generalized Additive Mixed Models (GAMMs; Wood, 2017) with two model specifications, one including random effects and the other incorporating random smooths, to assess the predictive capacity of these metrics while controlling for object proportion, visual saliency, and random factors such as participant and object position. Results demonstrate that all proposed metrics significantly predict visual attention (all p-values < 0.0001), with distinct predictive patterns revealing different cognitive mechanisms. Vision-based metrics showed clear relationships with attention: DINOV2-based obj_image_vissim exhibited the strongest effects (ΔAIC = -311 for duration, -301 for fixation count with random smooths), demonstrating monotonic patterns where objects visually similar to the scene context consistently attract more attention. CLIP-based metrics performed moderately (obj_image_vissim: ΔAIC = -208 for duration, -211 for fixation;

objs_vissim: ΔAIC = -143 to -156), while overall_vissim showed weaker contributions (ΔAIC = -125 to -140 for CLIP). Language-based metrics revealed complex, non-linear patterns: overall_semsim demonstrated strong performance (ΔAIC = -294 for duration, -306 for fixation with random smooths), with U-shaped relationships suggesting that both highly semantically relevant and semantically anomalous objects capture attention through different mechanisms. Words_semsim showed robust effects (ΔAIC = -252 for duration, -296 for fixation), while sent_semsim exhibited moderate performance (ΔAIC = -155 to -258), as shown in Fig.2. Notably, the combined metric (total_vissem_sim) consistently outperformed all individual metrics (ΔAIC = -382 for duration, -398 for fixation with random smooths), representing improvements of 20-30% over the best single-modality metrics and demonstrating that visual and linguistic information synergistically shape attention allocation. Temporal analysis revealed that vision-based metrics show stronger effects during early fixation periods (<200ms), consistent with rapid bottom-up feature extraction, while language-based metrics maintain consistent strength across temporal windows, reflecting sustained top-down semantic integration. The combined model's superior performance across both early and late periods underscores the importance of multimodal integration throughout visual processing.

These findings advance our understanding of visual cognition by demonstrating that human attention integrates information across multiple representational formats. The superior performance of combined metrics suggests that cognitive models must account for vision-language interactions rather than treating modalities independently. These results align with theories emphasizing multimodal processing (Holler & Levinson, 2019; Benetti et al., 2023) and illuminate the interplay between top-down semantic guidance and bottom-up perceptual salience (Gilbert & Li, 2013; Dijkstra et al., 2017). While our results are based on DINOV2 and CLIP embeddings, the framework is model-agnostic and applicable to emerging architectures. The computational complexity analysis demonstrates scalability ($O(k^2 \cdot d + |C| \cdot d)$ for k objects and d-dimensional embeddings, ~0.3 seconds per image), making the approach practical for large-scale applications. This work has practical implications for AI systems requiring human-like attention mechanisms, including image captioning, visual question answering, and assistive technologies for populations with attentional differences. By bridging computational modeling with empirical eye-tracking data, this research contributes to cognitive science, human-computer interaction, and the development of cognitively-inspired artificial intelligence systems that can more effectively model and predict human visual behavior.

References:

Benetti, S., Ferrari, A., & Pavani, F. (2023). Multimodal processing in face-to-face interactions: A bridging link between psycholinguistics and sensory neuroscience. *Frontiers in Human Neuroscience*, 17, 1108354.

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2017). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, 21(9), 652-665.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5), 350-363.

Hayes, T. R., & Henderson, J. M. (2021). Looking at the scene: Semantic similarity and visual attention. *Psychological Science*.

He, S., et al. (2019). Human attention in image captioning dataset. *CVF*.

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8), 639-652.

Hwang, A. D., et al. (2011). Semantic guidance of eye movements. *Vision Research*, 51(10), 1192–1205.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *EMNLP*.

Sun, K., & Liu, H. (2025). Attention-aware semantic relevance predicting Chinese sentence reading. *Cognition*, 255, 105991.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC.

Wu, B., et al. (2020). Visual transformers: Token-based image representation. *ECCV*.

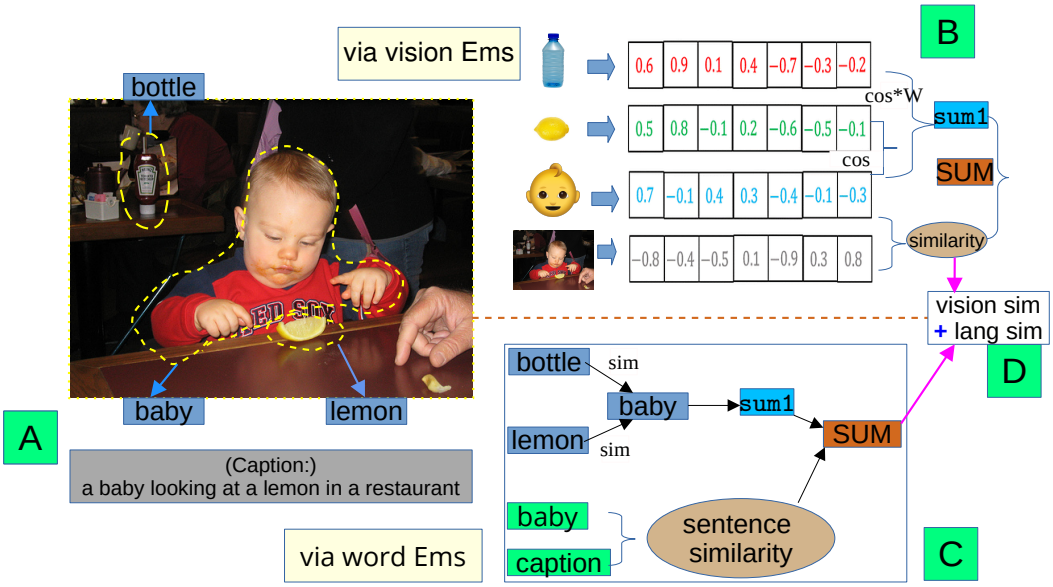

Fig. 1: **The computational method for contextual semantic relevance**. Panel A shows the input image with labeled objects (bottle, baby, lemon) and corresponding caption. Panel B illustrates vision-based metric computation: embeddings from Visual Transformer are compared via cosine similarity, then weighted and summed to produce metrics quantifying object-to-scene visual relationships. Panel C depicts language-based metric computation: word embeddings from Sentence Transformer measure semantic similarity between object labels and caption elements. Panel D shows the integration strategy: vision-based and language-based similarities are combined (weighted sum) to create the total_vissem_sim metric that captures multimodal semantic relevance. "Ems" = embeddings.

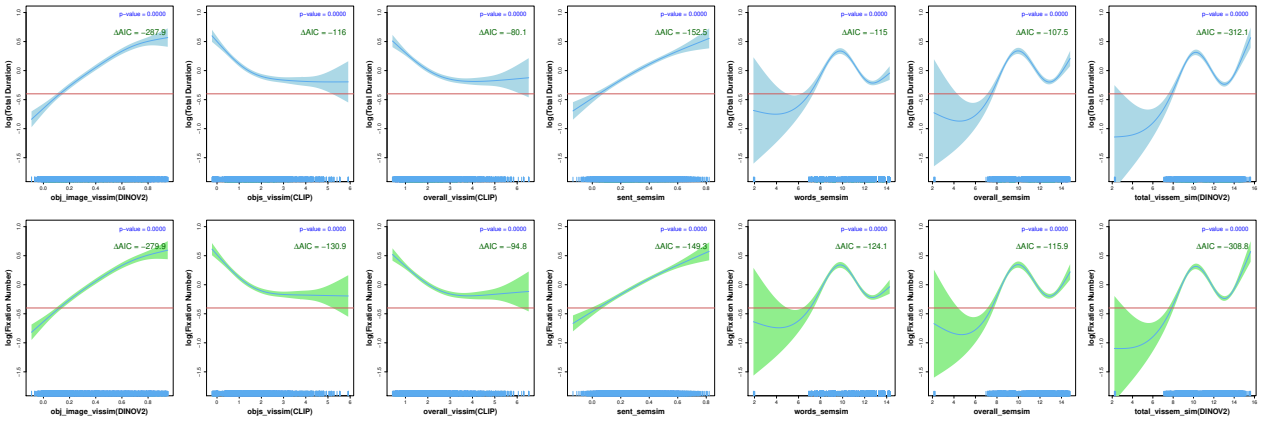

Fig. 2: **Partial effects of contextual semantic relevance metrics on visual attention measures using GAMMs with random effects**. The figure displays smooth functions showing how each metric predicts fixation behavior while controlling for object proportion, saliency, participant variation, and object position. The top row (blue/cyan) shows effects on log-transformed total fixation duration, while the bottom row (green) shows effects on log-transformed fixation count. From left to right, the columns present: (1-3) vision-based metrics using DINOV2 (obj_image_vissim) and CLIP (objs_vissim, overall_vissim); (4-6) language-based metrics (sent_semsim, words_semsim, overall_semsim); and (7) the combined multimodal metric (total_vissem_sim). Vision-based metrics demonstrate clear monotonic trends, with DINOV2's obj_image_vissim showing the strongest predictive power (ΔAIC = -287.9 for duration, -279.9 for fixation count). Language-based metrics exhibit more complex non-linear patterns: sent_semsim shows a strong positive relationship, words_semsim displays an initial dip followed by sharp increases and subsequent fluctuations, and overall_semsim shows moderate complexity. The combined metric achieves superior performance (ΔAIC = -312.1 for duration, -308.8 for fixation count), outperforming all individual metrics by 8-10% and demonstrating the synergistic benefit of integrating visual and linguistic information. Shaded regions represent 95% confidence intervals around the smooth functions. All effects are highly significant ($p < 0.0001$). More negative ΔAIC values indicate better model fit relative to the baseline model containing only control predictors and random effects. The x-axes show the range of metric values in the dataset, while *y*-axes represent the partial effect on log-transformed fixation measures (centered at zero).