Improving Vision-LLMs with Human Cognitive Signals

Introduction and Motivation

Large Language Models (LLMs) have made remarkable progress in recent years exceling across Natural Language Processing (NLP) tasks. Multimodal Large Language Models (MLLMs) build on them by extending beyond text to heterogeneous data. When inputs include images, these models are referred to as Vision Large Language Models (VLLMs). The core challenges that affect LLMs persist on VLLMs and are often amplified by the added visual modality. These challenges include alignment with human intentions, hallucinations, seamless multimodal integration, and growing data bottlenecks [1]. Current VLLMs are often poorly aligned with users' attentional focus and intentions, leading to imprecise or unhelpful responses in everyday scenarios where hallucinations are especially acute [2]. Recent work reports that many visual hallucinations are closely tied to *attention sinks* in the self-attention over image tokens [3]. These findings have motivated a new line of methods that explicitly analyse and manipulate attention and thereby reduce hallucination [3].

On the other hand, cognitive signals, particularly Eye-tracking (ET), have shown promise in addressing some of these challenges. They can enrich datasets with user-derived behavioural information [4], and recent work has begun to explore its potential in human alignment [4] since attention patterns can predict explicit human feedback [5]. The integration of vision and language in VLLMs has enabled additional opportunities to include visual attention on visual input with recent integration as shown by Yan et al. [2]. However, acquiring cognitive data in laboratory settings remains challenging due to the need for costly, specialised equipment and privacy concerns [4]. To address these limitations, generative models capable of producing synthetic data at scale have emerged, expanding the range of applications [2, 5].

Proposed approach

Given the potential of these cognitive signals and the recent emergence of datasets and studies linking them to human alignment, we propose to acquire cognitive signals with the Tobii Pro Fusion device[1] during the annotation process of a VLLMs alignment dataset for the first time. Because the effectiveness of LLMs often hinges on large training datasets, and because gaze is typically unavailable at inference, many approaches rely on cognitive prediction models [4]. We posit that leveraging human attention in VLLMs will likely require models that can accurately predict human saliency in this setting; therefore, we will fine-tune models such as UniAR [5] on the proposed dataset, and evaluate what advantages this provides over synthetic ET, such as the one proposed in Yan et al. [2], evaluating with saliency metrics like NSS, AUC and SIM.

We consider particularly relevant the line of work comparing human and model attention at both text and visual levels. Given the rise of attention-modification methods for reducing hallucinations [2], we will use this dataset and the resulting model to study both human-alignment techniques (e.g., Direct Preference Optimization (DPO)) and attention-editing strategies, evaluated on alignment and hallucination benchmarks.

References

[1] Stephen Casper, Xander Davies, Claudia Shi, and et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *TMLR*, July 2023.

[2] Kun Yan, Lei Ji, Zeyu Wang, and et al. Voila-A: Aligning Vision-Language Models with User's Gaze Attention. In *NeurIPS*, December 2024.

[3] Seil Kang, Jinyeong Kim, Junhyeok Kim, and et al. See What You Are Told: Visual Attention Sink in Large Multimodal Models. In *ICLR*, April 2025.

[4] Ángela López-Cardona and Carlos et al. Segura. Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models. In *ICLR*, April 2025.

[5] Peizhao Li, Junfeng He, Gang Li, and et al. UniAR: A Unified model for predicting human Attention and Responses on visual content. In *Advances in NeurIPS*, December 2024.

[1] `https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion`