Multimodal Large Language Models “Foresee” Objects Based on Verb Information But Not Gender

Anonymous ACL submission

Abstract

This study employs the classical psycholinguistics paradigm, the visual world eye-tracking paradigm (VWP), to explore the predictive capabilities of multimodal large language models (MLLMs) and compare them with human anticipatory gaze behaviors. Specifically, we examine the attention weight distributions of LLAVA when presented with visual displays and sentences containing verb and gender cues. Our findings reveal that LLAVA, like humans, can predictively attend to objects relevant to verbs, but fails to demonstrate gender-based anticipatory attention. Layer-wise analysis indicates that the middle layers of the model are more related to predictive attention than the early or late layers. This study is pioneering in applying psycholinguistic paradigms to compare the multimodal predictive attention of humans and MLLMs, revealing both similarities and differences between them.

1 Introduction

Recent psycholinguistic research has shown that human language processing involves multimodal predictions, especially between language and vision (e.g., Altmann & Kamide, 1999; see Huettig et al., 2011, for a review). For instance, numerous visual world paradigm (VWP) studies have demonstrated that when people hear an utterance, they predict upcoming mentions, which direct their looks to the visual objects. For example, in Corps et al. (2022), participants heard a sentence featuring either male or female characters and looked at the visual display of four objects at the same time (Figure 1). They found that: (1) participants used verb semantics to predict upcoming mentions (e. g., looking at wearable objects such as a tie or dress at hearing *Tonight, James/Kate will wear ...*); (2) they further used the gender of the subject to refine their prediction (e.g., more looks to a tie than a dress following *James*, and more looks to a dress than a tie following *Kate*).

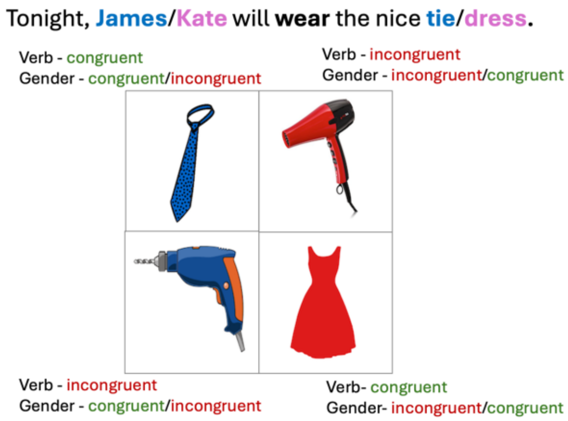

Figure 1: Sample visual display adapted from Corps et al. (2022)

The finding that humans use linguistic (verb and gender) information to make predictive fixations of a visual scene led us to ask whether multimodal large language models (MLLMs) exhibit similar cross-modal predictive behaviors. Previous studies have found parallels between model attention model attention (measured by attention weights) and human attention (measured by eye tracking movements) in text reading (Gao et al., 2023; Kewenig et al., 2024; Sood et al., 2020). Kewenig et al. (2024) recently provided tentative evidence that multimodal models like CLIP (Radford et al., 2021) may also resemble human predictive visual attention in video viewing. But to our best knowledge, there is not yet research on whether MLLMs also resemble humans in linguistically-guided predictive visual attention.

The current study employs the widely adopted VWP in psycholinguistics to investigate whether LLAVA (Liu et al., 2023), an open-source MLLM, shows similar linguistically-guided predictive visual attention as humans. By analyzing the model's attention weight distribution on the task used by Corps et al. (2022), we found that MLLM can predictively attend to relevant objects based on verb information, similar to humans, but not gender information. In addition, layer-wise analysis shows that the middle layers of MLLM are primarily responsible for the predictions. These findings indicate both similarities and differences between model and human behavior in multimodal predictions.

2 Methods

2.1 Design and materials

Following Corps et al. (2022), we used 28 pairs of sentences featuring either male or female characters (e.g., *Tonight, James/Kate will wear the nice tie/dress*), each with a visual display of four objects (Figure 1). We tested whether an MLLM can predictively attend to a visual object according to whether the object is verb-congruent (e.g., dress and tie for the verb *wear*) or verb-incongruent (e.g., drill and hairdryer), and whether this prediction (if any) is further modulated by the object's congruency with the gender of the sentential subject (e.g., for *James*, tie and drill are gender-congruent and dress and hairdryer are gender-incongruent; for *Kate*, the conditions are reversed). The object images are 200×200 pixels, with their locations counterbalanced across items.

2.2 Model

We utilized LLAVA 1.5 (7B parameters, Liu et al., 2023), a transformer-based MLLM that encodes images using CLIP's vision encoder and maps them into the linguistic embedding space of Vicuna, allowing cross-modal attention to be computed. This model was chosen for its open-source availability and its state-of-the-art performance on 11 benchmarks (Liu et al., 2023).

2.3 Pre-tests

We first conducted three pre-tests to explore if LLAVA can recognize the basic information in sentences and pictures as humans do.

(1) Name gender detection. To investigate if the model can distinguish gender based on names (*James* vs. *Kate*), we asked the model to continue a sentence preamble (e.g., *Although James/Kate was sick…)* and calculated the proportions of female (*she/her/hers*) or male pronouns (*he/his*) used in the continuations following Cai et al. (2023, experiment 2). We found that all sentences with *James* were continued with male pronouns and all sentences with *Kate* were continued with female pronouns. This indicates that the model can perfectly distinguish between typical male and female names in sentences.

(2) Object gender evaluation. To assess whether the model can identify pictured objects as stereotypically male (e.g., tie, drill) or female (e.g., dress, hairdryer), we asked the model to evaluate the masculinity or femininity of each object on a 5-point Likert scale and calculated the “femininity score” of each object where 1 represents strongly masculine and 5 represents strongly feminine. The results show that the femininity score of female objects is significantly higher than that of male objects (3.13 vs. 2.67; $t(5641.6) = 11.202$, $p < .001$), indicating that the model can identify the gender of the objects.

(3) Multimodal sentence completion. To examine whether the model can complete the sentence with verb-and-gender-congruent nouns in a multimodal setting, we removed the final noun from the sentence and asked the model to complete the missing linguistic materials according to the sentence's corresponding visual display. As shown in Figure 4 in Appendix A, the model produced more verb-congruent completions than incongruent ones (83.77 vs. 12.52; $t(109.29) = -11.844$, $p < .001$), and also more gender-congruent completions than incongruent ones (64.61 vs. 29.52; $t(109.83) = -4.2849$, $p < .001$). This indicates that the model can predict verb-and-gender-congruent nouns in a multimodal sentence completion task.

2.4 Procedure

To simulate human incremental sentence comprehension, we presented the sentence in an unfolding fashion, ending first with the name (e.g., *Tonight, James/Kate*), then with the verb (e.g., *Tonight, James/Kate will wear*), then with the pre-noun adjective (e.g., *Tonight, James/Kate will wear the nice*), and finally the whole sentence ending with the target noun (e.g., *Tonight, James/Kate will wear the nice tie/dress*). Each text presentation was accompanied by the same visual display of four objects. We used the prompt: "Please read carefully and look at the objects in the picture," which mirrors the instructions given to human participants, ensuring that the model's task closely parallels the one performed by human subjects.

3 Analyses and results

3.1 Analysis

We extracted the max-pooled attention weights of each layer mapping from the last word (name, verb, pre-noun adjective, or target noun) of each sentence segment to the four images in the visual display. Following Manning et al. (2020), if the last word had multiple tokens, we combined the weights across the tokens. We then calculated the proportion of attention allocated to each object relative to the total attention across all four objects, similar to fixation proportions in VWP studies (e.g., Corps et al., 2022).

For statistical analysis, we used linear mixed-effect models, with verb congruency and gender congruency as interacting predictors, and with layer and item as random effects. Following Matuschek et al. (2017), we used forward model comparison with an alpha level of 0.2 to determine whether a random slope should be included in the final model.

3.2 Results

3.2.1 Main results of the whole model

Figure 2 (top panel) shows the attention proportions to four objects across sentence segments. Initially, when the name was read, LLAVA showed no preference for gender-congruent objects ($\beta = 0.001$, $SE = 0.002$, $t = 0.326$, p = 0.744), suggesting that the model did not associate specific objects with the gendered name in the absence of further contextual information.

As the sentence unfolded to the verb (e.g., *wear*), there is a significant preference for verb-congruent objects (e.g., tie and dress) over incongruent ones (e.g., drill and hairdryer; $\beta = 0.009$, $SE = 0.002$, $t = 4.168$, $p < .001$), indicating that LLAVA can use verb semantics to direct attention similar to humans. Nevertheless, there was still no effect of gender congruency ($\beta = 0.0004$, $SE = 0.002$, $t = 0.187$, $p = .852$), suggesting that the model still does not preferentially attend to gender-congruent objects at this stage.

As the model received more input (e.g., *Tonight, James/Kate will wear the nice ...*), the difference between verb-congruent and verb-incongruent objects continued to grow ($\beta = 0.035$, $SE = 0.004$, $t = 8.599$, $p < .001$) and the absence of a gender congruency effect persisted ($\beta = -0.002$, $SE = 0.003$, $t = - 0.862$, $p = .389$).

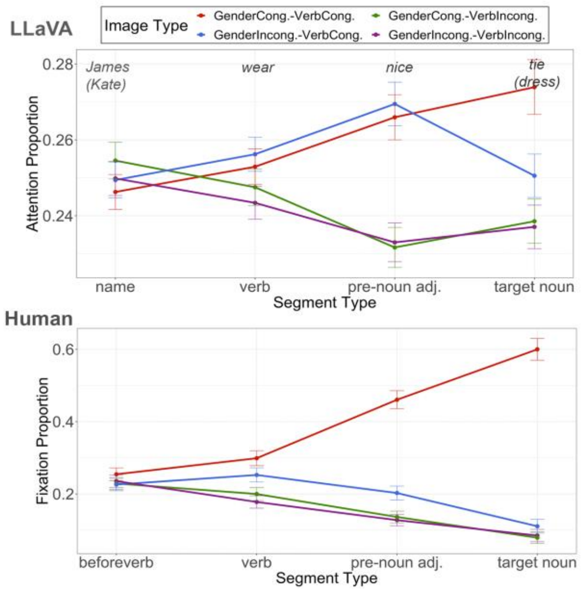

Figure 2: Compare attention proportion of LLAVA (top panel) and fixation proportion of humans (bottom panel; data from Corps et al., 2022)

Finally, when the sentence was fully presented (e.g., *Tonight, James/Kate would like to wear the nice tie/dress)*, the pattern remained unchanged, with a significant effect of verb congruency (β = 0.024, SE = 0.003, t = 9.491, p < .001), but no evidence of a gender congruency effect ($\beta = 0.012$, $SE = 0.019$, $t = 0.640$, $p = .527$).

We compared our model attention and human attention (fixation proportion) in Corps et al. (2022) (see Appendix B for detailed methods). As shown in Figure 2, there are both similarities and differences in anticipatory processing patterns. Overall, the patterns between humans and the model are similar ($r = 0.11$, $p = .018$). However, this similarity is primarily driven by the verb factor ($r = 0.25$, $p < .001$), and not by the gender factor ($r = -0.009$, $p = .896$; see Figure 6 in Appendix B). This is because human participants not only predictively attended to verb-relevant objects ($\beta = 0.087$, $SE = 0.010$, $t = 9.112$, $p < .001$), but also gender-relevant objects ($\beta = 0.034$, $SE = 0.016$, $t = 2.147$, $p = 0.040$) as soon as they heard the verb.

3.2.2 Results of layer-wise analysis

In addition to analyzing the overall behavior of the model across all layers, we conducted a more fine-grained, layer-wise analysis to identify the layers that were primarily responsible for the verb-based predictive visual attention in LLAVA. As shown in Figure 3, our results indicate that the middle layers of the model play a crucial role in generating visual predictions based on verb information.

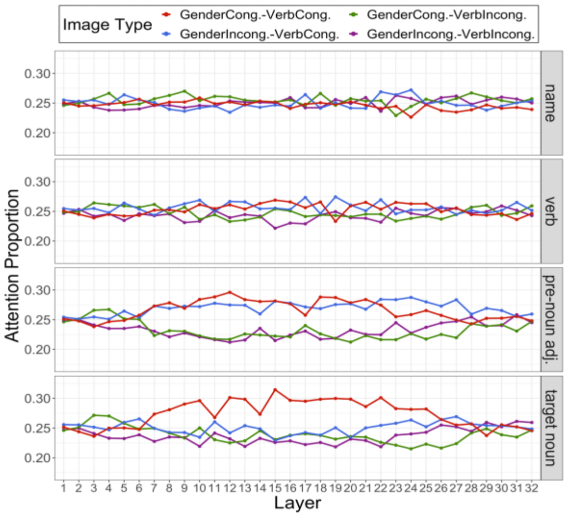

Figure 3: Attention results by layers

During the verb segment of the sentence (e.g., *James/Kate will wear*), we found a significant main effect of verb (*ps* < .05) in layers 10, 12, and 17 (see the second panel in Figure 3). As the sentence unfolds (e.g., *James/Kate will wear the nice*), the main effect of verb becomes more widespread, occurring in layers 7 through 26 (see the third panel in Figure 3). This indicates that a larger portion of the model's architecture is engaged in verb-based predictions as more linguistic context becomes available.

4 Discussion

This study uses the VWP to investigate the predictive capabilities of MLLMs like LLAVA. The findings reveal that the model exhibits human-like behavior in using verb information to predict the upcoming object in a visual display. This aligns with previous research demonstrating that both humans and models can utilize multimodal information to predictively attend to relevant features (Kewenig et al., 2024).

However, unlike humans, the model does not predictively attend to relevant objects based on gender information, consistent with the lack of gender bias in CLIP, which is the basis for LLAVA's vision encoder (Hall et al., 2024). Indeed, our pretest results showed that although the model can complete sentences with gender-congruent nouns, it does not do so to the same extent as it produces verb-congruent nouns (see Figure 4 in Appendix A). This uncertainty in behavioral responses is consistent with the lack of a gender effect in the attention weights.

The difference between the model and humans may be explained by the nature of the stimuli, as our study used cartoon-like images while MLLMs are mainly trained and evaluated on real-world objects (Liu et al., 2023; Thrush et al., 2022). To investigate this hypothesis, we replaced the cartoon-like objects with real-world ones. As shown in Figure 7 in Appendix C, we observed a main effect of gender in the verb segment (β = 0.009, *SE* = 0.002, t = 4.117, p < .001), suggesting that the model processes real-world objects in a more human-like way than cartoon objects. This is consistent with the idea that models lack the perceptual flexibility of humans, leading to lower performance in recognizing atypical objects (Zang et al., 2023).

The study also found that the middle layers play a significant role in multimodal predictions, aligning with previous studies showing that attention weights in middle layers better fit neural signals (Lamarre et al., 2022). However, the discrepancy with some studies showing that late layers correlate most significantly with human eye-tracking data (Kewenig et al., 2024) may be attributed to task differences: comprehension tasks (as in our and Lamarre et al.'s studies) require more high-level semantic processing in middle layers, while production tasks (as in Kewenig et al., 2024) focus more on low-level features of individual words in later layers. Further detailed experiments are needed to explore this hypothesis.

5 Conclusion

In conclusion, our study utilizes the VWP from psycholinguistics to probe whether MLLMs like LLAVA show similar multimodal predictive patterns to humans. We found that MLLMs can predictively attend to verb-relevant objects in visual displays similar to humans, but they do not show the same predictive attention for gender-relevant objects. These predictive behaviors are predominantly driven by the middle layers of the model.

Limitations

One key limitation of this study is that we only investigated one model — LLAVA-1.5 7B — and conducted a thorough comparison between its attention weights and human eye movements. With more MLLMs being released (see Yin et al., 2024 for a comprehensive review), it is crucial to compare different models horizontally to understand the key factors contributing to their differences and similarities with human cognition.

In addition, caution is needed when comparing human and model attention. Although both use the term "attention," they may refer to different underlying mechanisms. For instance, model attention is more evenly dispersed, while human attention tends to be focused (Kewenig et al., 2024; also see Figure 2). More detailed studies are needed to explore the similarities and differences between model attention and human attention.

Ethical considerations

The authors declare no competing interests. The stimuli used are provided by the first author of Corps et al. (2022) via email. The human eye-tracking data used is publicly available (https://osf.io/nkud5/) and does not contain personal information about the subjects. The usage scenario of the model LLAVA conforms to its licensing terms. As this work focuses on comparing the multimodal predictions of models and humans, its potential negative impacts on society seem to be minimal.

References

Gerry TM Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Zhenguang G. Cai, David A. Haslett, Xufeng Duan, Shuqi Wang, and Martin J. Pickering. 2023. Does ChatGPT resemble humans in language use? *arXiv preprint* arXiv:2303.08014 [cs]. Version 2.

Ruth E. Corps, Charlotte Brooke, and Martin J. Pickering. 2022. Prediction involves two stages: Evidence from visual-world eye-tracking. *Journal of Memory and Language*, 122:104298.

Changjiang Gao, Shujian Huang, Jixing Li, and Jiajun Chen. 2023. Roles of Scaling and Instruction Tuning in Language Perception: Model vs. Human Attention. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13042–13055, Singapore. Association for Computational Linguistics.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

Falk Huettig, Joost Rommers, and Antje S. Meyer. 2011. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2):151–171.

Yuki Kamide, Gerry TM Altmann, and Sarah L. Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, 49(1):133–156.

Viktor Kewenig, Andrew Lampinen, Samuel A. Nastase, Christopher Edwards, Quitterie Lacome DEstalenx, Akilles Rechardt, Jeremy I. Skipper, and Gabriella Vigliocco. 2024. Multimodality and Attention Increase Alignment in Natural Language Prediction Between Humans and Computational Models. *arXiv preprint* arXiv:2308.06035 [cs]. Version 3.

Mathis Lamarre, Catherine Chen, and Fatma Deniz. 2022. Attention weights accurately predict language representations in the brain. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4513–4529. Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved Baselines with Visual Instruction Tuning. *arXiv preprint* arXiv: 2310.03744 [cs].

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Hannes Matuschek, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. Balancing Type I error and power in linear mixed models. *Journal of memory and language*, 94:305–315.

Martin J. Pickering and Chiara Gambi. 2018. Predicting while comprehending language: A theory and review. *Psychological Bulletin*, 144(10):1002–1044. 113.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish

Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021.Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, PMLR 139*.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension. In Raquel Fernández and Tal Linzen, editors, *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A Survey on Multimodal Large Language Models. *arXiv preprint* arXiv:2306.13549 [cs].

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2023. Contextual Object Detection with Multimodal Large Language Models. *arXiv preprint* arXiv:2305.18279 [cs].

A Multimodal sentence completion task

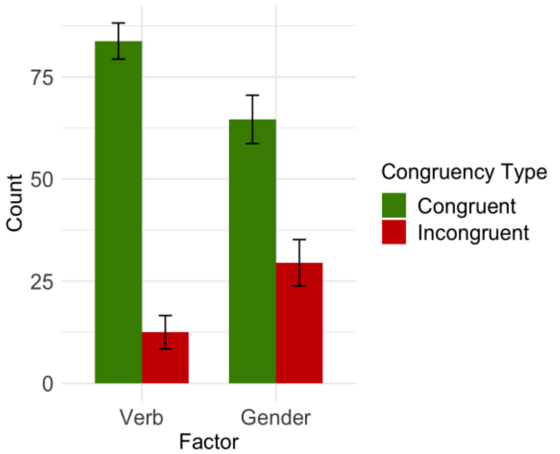

Figure 4: Results of sentence completion task

B Compare with human data

Since eye movement data in Corps et al. (2022, accessible at https://osf.io/nkud5) were analyzed at 50ms intervals, we need to transform the data into four segments to align with the model data. According to the R scripts available at https://osf.io/nkud5/, the four segments are defined as follows:

- Before verb: < 0ms (before verb onset)
- Verb: 0-350ms (from verb onset to verb offset)
- Pre-noun adjective: 350-850ms (from verb offset to target onset)
- Target: >850ms (after target onset)

Within each segment, we aggregated fixation points and calculated the fixation proportion of each object. These aggregated data were then used for further analysis and plotting. This transformation ensures the human data is comparable with the model data. From Figure 5, we can observe that the reshaped data exhibit a similar pattern to the original data.

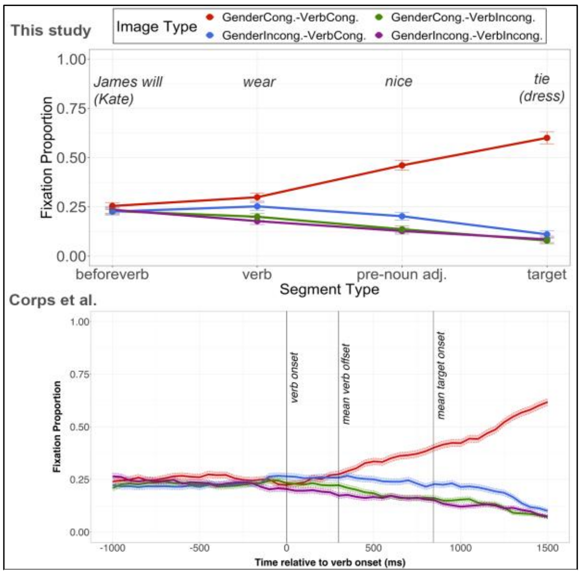

Figure 5: Compare plots of humans in our study (top panel) and Corps et al. (2022, bottom panel)

We also calculated the Pearson correlations between human and model data. We grouped the data by both gender and verb factors, only by verb factors and only by gender factors and then calculated the correlations respectively. The results are shown in Figure 6.

C Attention to real-world objects

For each object picture in the stimuli, we search for a similar picture in Google Images (the same source as Corps et al., 2022) but with a real-world object. We replaced each object picture with the new real-world one and conducted the experiment again. The results are shown as in Figure 7.

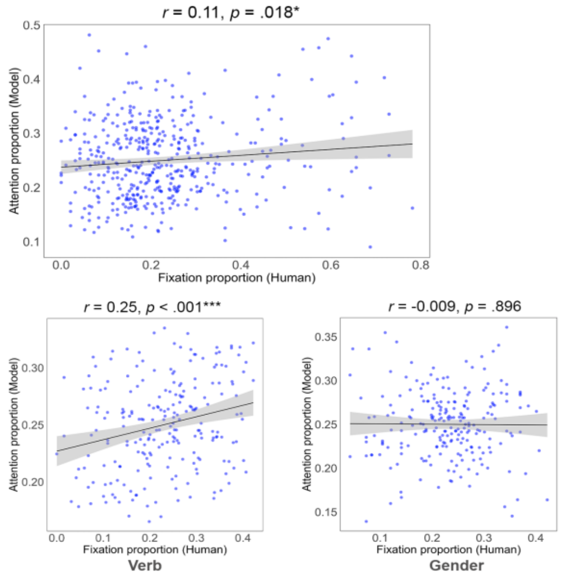

Figure 6: Correlation between the model and humans when considering both gender and verb factors (top), only verb factor (left-bottom), and only gender factor (right-bottom)

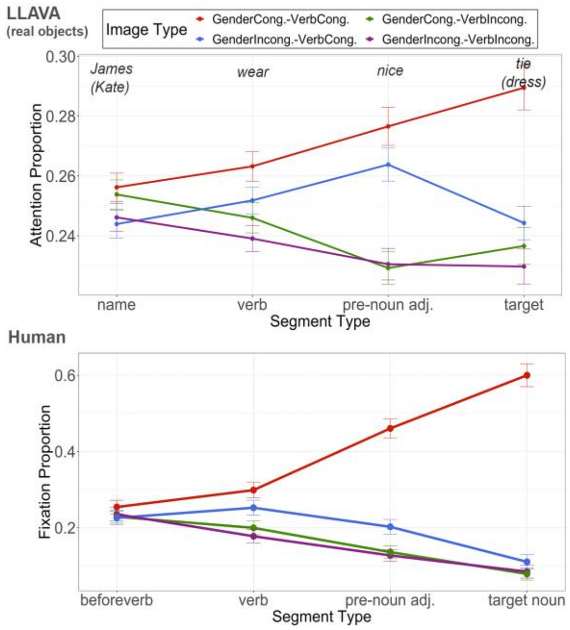

Figure 7: Compare model attention proportions using real-world stimuli in LLAVA (top) and fixation proportions of humans (bottom)