Image-conditioned human language comprehension and psychometric benchmarking of visual language models

Subha Nawer Pushpita
Masachusetts Institute of Technology
snpushpi@mit.edu

Roger Levy
Massachusetts Institute of Technology
rplevy@mit.edu

Abstract

Large language model (LLM)s' next-word predictions have shown impressive performance in capturing human expectations during real-time language comprehension. This finding has enabled a line of research on psychometric benchmarking of LLMs against human language-comprehension data in order to reverse-engineer humans' linguistic subjective probability distributions and representations. However, to date this work has exclusively involved unimodal (language-only) comprehension data, whereas much human language use takes place in rich multimodal contexts. Here we extend psychometric benchmarking to visual language models (VLMs). We develop a novel experimental paradigm, *Image-Conditioned Maze Reading*, in which participants first view an image and then read a text describing an image within the Maze paradigm, yielding word-by-word reaction-time measures with high signal-to-noise ratio and good localization of expectation-driven language processing effects. We find a large facilitatory effect of correct image context on language comprehension, not only for words such as concrete nouns that are directly grounded in the image but even for ungrounded words in the image descriptions. Furthermore, we find that VLM surprisal captures most to all of this effect. We used these findings to benchmark a range of VLMs, showing that models with lower perplexity generally have better psychometric performance, but that among the best VLMs tested perplexity and psychometric performance dissociate. Overall, our work offers new possibilities for connecting psycholinguistics with multimodal LLMs for both scientific and engineering goals.

1 Introduction

Human language comprehension is highly incremental. Our minds integrate linguistic input with context very rapidly: words within sentences, and even phonemes or letters within spoken or written words, to update our understanding of linguistic input (Tanenhaus et al., 1995; Rayner, 1998). This process involves the rapid update of expectations about the interpretation of what has already been said and predictions about what might be said next. These predictions affect how we process the language we encounter, helping us to recognize and correct errors (Marslen-Wilson, 1975; Levy, 2008b) and to analyze input more rapidly.

The fundamental operation of large language models (LLMs) is similar: LLMs put probability distributions over the next tokens given the preceding context. This convergence has made it natural to compare LLM distributions with human linguistic behavior. In unimodal language processing, LLM predictions have been shown to align fairly well with those generated by humans in the Cloze task (Goldstein et al., 2022). Furthermore, there is a linear relationship between the surprisal of a word in linguistic context (negative log-probability; (Hale, 2001; Levy, 2008a)) and how long comprehenders take to read it (Smith and Levy., 2013; Wilcox et al., 2023). These findings have generated interest in psychometric benchmarking of language models (LMs): comparing LMs in terms of how well their autoregressive probabilities predict human reading times or other types of linguistic behavior (Frank and Bod, 2011; Fossum and Levy, 2012; Goodkind and Bicknell, 2018; Oh and Schuler, 2023; Shain et al., 2024).

Psychometric benchmarking of LLMs has exclusively involved unimodal, language-only data and models. However, human language use generally involves a rich multimodal context. For this reason, there is growing interest in multimodal language models. The most advanced such type of model is vision-language models (VLMs), which relate visual content (most commonly static images) to linguistic content. For example, models like BLIP-2 (Li et al., 2023) can generate text associated with an image; to do this, it autoregressively places con-

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 447–457
November 15-16, 2024 ©2024 Association for Computational Linguistics

ditional probability distributions over next linguistic tokens given an image in context plus preceding linguistic context. However, evaluation techniques for VLMs are less developed than for unimodal LLMs, and we are aware of no work to date on psychometric benchmarking for VLMs.

Here we present a framework and experimental results on psychometric evaluation of visual language models using a novel yet simple psycholinguistic experimental paradigm. In an experimental trial, a participant first previews an image, then reads a sentence describing an image, with word-by-word reading times measured (Figure 1). The image may be the one that the sentence describes (the **Correct Image** condition), a different image that the sentence does not describe (the **Wrong Image** condition), or simply a black screen (the **No Image** condition). Intuitively, previewing the correct image should prepare the participant for the sentence description and facilitate them reading it more quickly and accurately. However, there are different forms that this facilitation could take, corresponding to different theoretical accounts of how visual context shapes language processing. Additionally, we can compare VLMs in terms of how well they capture how different image contexts influence the participant's reading behavior. We can thus use this experimental paradigm both to gain insight into the role of visual context in language processing in the human mind and to psychometrically benchmark visual language models. All the experiment codes, analysis, and datasets used in the project are made available at the linked repository. [1].

2 Related Work

2.1 Human vision and language processing

There is considerable psycholinguistic literature on the vision-language interface, with emphasis on visual context effects on spoken word recognition, syntactic disambiguation, and predictive processing.Much of this work uses the Visual World Paradigm (VWP), which investigates eye movements in visual scenes during spoken language understanding. Allopenna et al. (1998) and Dahan et al. (2001) used the VWP to demonstrate rapid, fine-grained effects of sub-word phonetic information on word-level interpretations, demonstrating incrementality of spoken language processing at the sub-word level. (Tanenhaus et al., 1995) used the VWP to demonstrate that the language processing system utilizes visual context to quickly interpret an ambiguous prepositional phrase, integrating lexical, syntactic, visual, and pragmatic reasoning. (Altmann and Kamide, 1999) showed how visual context aids predictive processing, supporting the idea that sentence comprehension involves anticipating the relationships between verbs, their syntactic components, and the real-world context they describe. For a broader review see Huettig et al. (2011).

2.2 Psychometric benchmarking of LLMs

It has long been known that words predictable in context are read faster (Ehrlich and Rayner, 1981) and elicit distinctive brain responses (Kutas and Hillyard, 1980; Kutas and Federmeier, 2011). Smith and Levy. (2013) found a linear relationship between n-gram word surprisal (negative log-probability) and reading time, a relationship that has held up with neural language models (Goodkind and Bicknell, 2018; Wilcox et al., 2023) and has been widely used to psychometrically benchmark LLMs (Oh and Schuler, 2023; Shain et al., 2024). There is also some evidence for a linear relationship between surprisal and the N400 ERP response (Heilbron et al., 2022, though see Szewczyk and Federmeier, 2022), and the best alignment of LM internal representations with brain activation patterns during language comprehension seems to be achieved by autoregressive LM architectures (Schrimpf et al., 2021; Caucheteux and King, 2022; Antonello et al., 2023). These results raise the prospect of reverse-engineering human subjective probabilities active during language processing through psychometric LLM benchmarking.

2.3 The Maze paradigm

Our experiment involves a simple adaptation of the Maze paradigm for studying word-by-word reading (Forster et al., 2009; Witzel et al., 2012; Boyce et al., 2020). In the Maze paradigm, experimental participants read a text passage through a sequence of two-alternative forced-choice tasks, one per word in the passage. Each word is coupled with an alternative distractor, one randomly assigned on the left and the other on the right, and the participant has to choose which word is correct (i.e., fits with the preceding linguistic context). The participant's reaction time (RT) and whether they chose the correct word are recorded. These reaction times

[1] https://github.com/snpushpi/Image-creates-linguistic-expectation

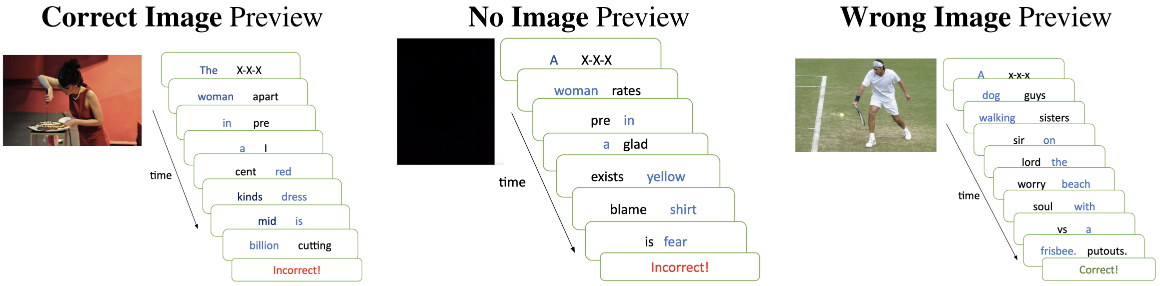

Figure 1: Schematic of image-description A-maze reading in each of the three experimental conditions. Participants first briefly view an image and then read a description by successively choosing the word fitting the preceding linguistic context and rejecting a foil word (example selections marked in blue). A mistake triggers an error message, and the participant moves on to the next trial sentence.

and accuracies carry information about the word's difficulty in a context that can be revealed through statistical analysis. The Maze paradigm has a number of methodological advantages: it is easily deployable over the web, it has a good signal-to-noise ratio, and processing difficulty is highly *localized*: that is, if a word is difficult for the comprehender, that difficulty shows up predominantly in RT and accuracy on that word, rather than "spilling over" to subsequent words as is often seen with other reading-time measurement techniques such as eye tracking or self-paced reading. Boyce and Levy (2023) showed that a linear relationship between surprisal and RT holds in the Maze paradigm as it does for other reading time-measuring paradigms.

3 Experimental Methodology

We developed an *Image-Conditioned Maze* experimental paradigm which is like the original Maze, but participants preview an image before reading each text passage. We chose 108 images and their corresponding descriptions from the validation split of Microsoft COCO (Lin et al., 2014). In each experimental trial, participants were first shown an image for 5 seconds, and then the image disappeared from the screen and they read an image description word by word in the Maze task. We generated distractor words using the A(uto)-Maze software of Boyce et al. (2020), which uses an LSTM RNN based model (Gulordava et al., 2018) to generate contextually unlikely words. Reaction time and response for each word choice (correct vs. distractor) were recorded. We recruited 69 US native English speaker participants (a quantity determined using power analysis based on a pilot study with a different set of images and descriptions) on Prolific, showed them some examples, and paid them 12$/hour for their participation. Each of them participated in 36 trials, 12 in each of the three conditions described before in figure (1), with trial order randomized for each participant. No participant saw the same image description twice.

In a separate study with different participants, we collected groundedness ratings for each word in each description in the context of the correct image associated with the description (Figure 2). We recruited 42 US native English speaker participants on Prolific for this study. Each sentence was rated by 7 participants on average. Participants used a slider to indicate how "present" each word was in the image, ranging from -10 (Not Present) to $+10$ (Surely Present).

4 Psycholinguistic hypotheses

Under wide circumstances, visual input automatically activates corresponding linguistic representations; a famous example is the Stroop effect, where a word naming one color but presented in another, such as blue, is difficult to say due to the interference between the words activated by the color versus orthographic information. We thus hypothesize that previewing the image will tend to activate at least some of the linguistic content in the image's description, so that reaction times will be faster and accuracy higher more quickly and accurately in the Correct Image condition than in the Wrong Image and No Image conditions. We also hypothesize that the Wrong Image condition may slow reaction times and reduce accuracy relative to the No Image condition, since the linguistic content that the image activates may conflict with the content in the subsequent text.

We distinguish between two versions of these hypotheses. One possibility is that activation of linguistic content may be restricted to content that is straightforwardly grounded in the image. For

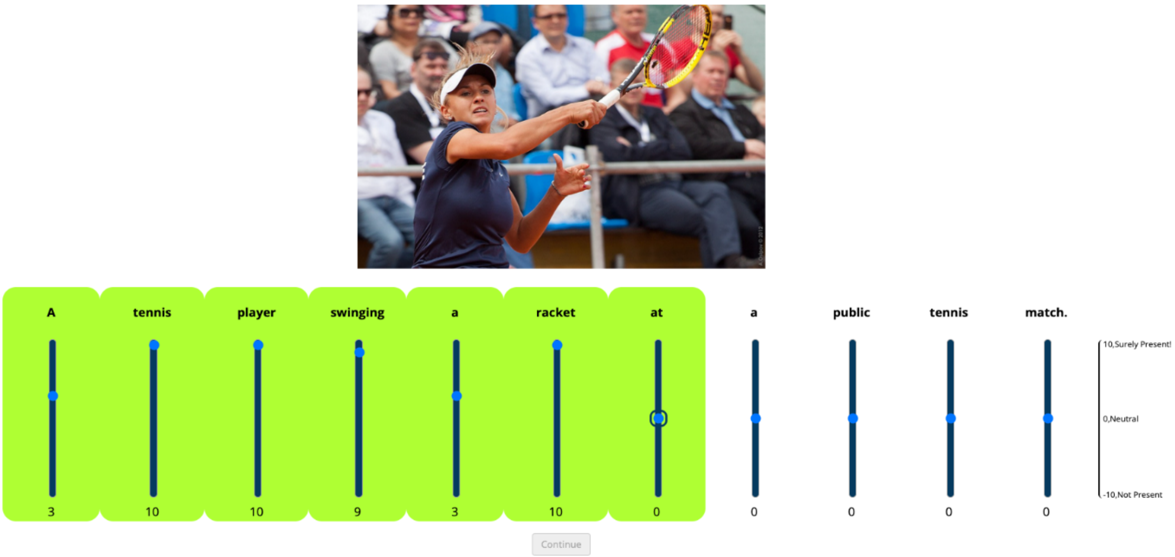

Figure 2: Example experiment page for a trial in the groundedness rating study. The circle indicates the slider the participant is currently manipulating. Once a participant chooses the vertical slider, the slider turns green. A participant must rate each word in the description to continue to the next trial. The scale on the right is a reminder of how the rating works.

example, in the Correct Image example of Figure 1, the words *woman*, *red*, and *dress* are straightforwardly grounded: the meaning of each word is prominent in the image without extensive reasoning or search for complex linguistic descriptions. In contrast, the rest of the words in that description are less straightforwardly grounded. Our **lexical-grounding hypothesis** is that linguistic facilitation or interference effects from the image will be limited to relatively straightforwardly grounded words. In cognitive terms, objects, properties, events, and states in the scene are visually identified, and the corresponding lemmas are activated so that when those lemmas are encountered in the image description, they are processed more effectively. We operationalize groundedness in two different ways: first as open-class (generally more grounded) versus closed-class (generally less grounded) parts of speech; second, through our grounding study as described in Section 3.

The second possibility, the **comprehensive-grounding hypothesis**, is that images evoke expectations over complete possible descriptions. This hypothesis predicts that facilitation or interference will affect all types of words in the sentence, regardless of part of speech or groundedness. A particularly strong version of the comprehensive-grounding hypothesis is that *all* facilitation and interference effects from the image will be mediated by this change in linguistic expectations. If this strong version of the hypothesis is correct, and if visual language models do a good job of capturing this shift in expectations, then visual language model surprisal should fully account for the effect of experimental conditions in the human behavioral data in our experiment.

5 Modelling Approach

We created a set of predictor variables including Condition_ID, frequency, word length, groundedness, open vs. closed part of speech, and surprisals from six Transformer-based LLMs: four visual language models with a variety of objectives regarding language-vision alignment (BLIP2, Li et al., 2023; KOSMOS2, Peng et al., 2023; LLAVA-7b, Liu et al., 2023; and IDEFICS-9b, Laurençon et al., 2024) and two language only models (GPT2, Radford et al., 2019; and LLAMA2 Touvron et al., 2023). Condition_ID indicates whether a certain image description was seen in Correct, Wrong, or No Image condition, which could be extracted from the experiment setup on IBEXZehr and Schwarz, 2018. For length, we used the length in characters excluding end punctuation. We obtain word frequencies from SUBTLEX_US (Brysbaert and New, 2009); for the words not in the database, we use the minimum frequency of any word in that database. Groundedness comes from our norming study. For open versus closed class part of speech, we ran the Stanford POS tagger on our image descriptions and considered all nouns, adjectives, adverbs, and non-auxiliary verbs, as open-class, and the rest as

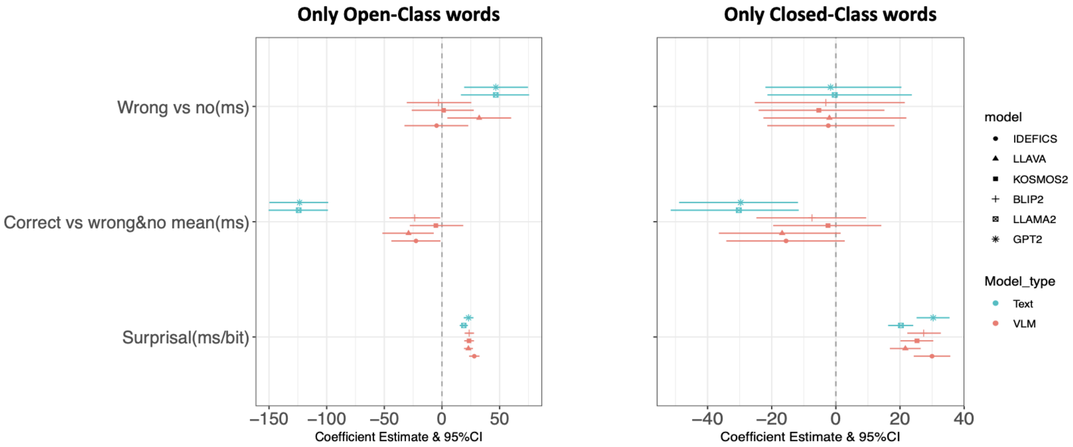

Figure 3: Coefficent Estimates and 95% CI of the fixed effects with theoretical interests for models fitted with open and closed class respectively. Condition_ID was Helmert encoded making comparisons between wrong vs no and correct vs wrong and no mean

closed-class. Surprisal does not vary across conditions for LLMs, but does so for VLMs for image conditioning. (Note that for the No Image condition, we used a black screen as the image, and additionally added "Ignore the image context" as a prompt preceding the description.) Using these predictors, for both testing our psycholinguistic hypotheses and psychometric benchmarking, we fitted mixed effects regression models to predict the reading time data that we collected, using the `brms` and `lmer` package in `R`. These models give us estimates and statistical significance of coefficients for all the predictor variables, which we can later analyze to distinguish between psycholinguistic hypotheses. For psychometric benchmarking, we fitted many models, each only varying at the kind of surprisal estimate it's using. For each fitted model, we then analyze the likelihood of the ground truth reading time data.

5.1 Regression predictor encoding

Unless otherwise specified, we used Helmert coding for Condition_ID, set up so that one coefficient encodes the **wrong** and **no** difference and another coefficient encodes the difference between **correct** and **(wrong** and **no)** mean. We used sum-encoding for open vs. closed part of speech (POS). Unless the model is condition specific, in which case Condition_ID can't be used as a predictor, we also assumed an interaction between Condition_ID and groundedness and Condition_ID and POS. Assuming this interaction makes sense since one would intuitively expect that one reads words in the correct condition even faster especially when the words are more highly grounded. For all the models, we use the maximal random effects structure justified by the design, so we have included correlated by-subject, by-sentence, by-word, and by-wordtoken random slopes for Condition_ID, the fixed effect of our primary theoretical interest. An example of a mixed effect model fitted for reading time prediction using data from all conditions and parts of speech(open vs. closed) is the following - `RT ∼ Condition_ID.helm*POS + surprisal + Frequency + Length + (Condition_ID.helm*POS + surprisal | Subject_ID)+ (Condition_ID.helm | Group) + (Condition_ID.helm | WordToken) + (Condition_ID.helm | Word)`.

6 Results

6.1 Reading Time Prediction

Consider figure (3), which plots the coefficient estimates and 95% confidence interval of the effects of theoretical interests from the model fitted with equation `RT ∼ Condition_ID.helm + surprisal + Frequency + Length + (Condition_ID.helm + surprisal | Subject_ID)+ (Condition_ID.helm | Group) + (Condition_ID.helm | WordToken) + (Condition_ID.helm | Word)`, individually for open and closed class words. Now note the second rows in both panels for models fitted with text-based surprisals(indicated in light blue in the figure). For the left panel, the second row is saying that on average people need 125 ms less to read an open class word in the correct condition compared to other conditions. Similarly, for the right panel, the second row indicates that on average people need 30ms less to read a closed class word in the correct condition compared to other conditions. So there is a very significant facilitation for both open and closed-class words when people get a preview of the relevant image compared to when they don't. This evidence strongly suggests that people's facilitation of reading image descriptions after having a relevant visual preview can be explained by **Comprehensive Grounding Hypothesis** and not by **Lexical Grounding Hypothesis**. Note that we want to consider only the text surprisal fitted models' condition-related effects to distinguish between lexical and comprehensive grounding hypotheses. It is because in this scenario the only image-related information we want to use for RT prediction should be through Condition_ID/POS levels. In both panels of Figure (3), we can see that the impact of condition ID-related effects is noticeably smaller—or even non-existent—in VLM surprisal-fitted models compared to text surprisal-fitted models. However, the overall effect of surprisal itself is quite similar across both types of models. To gain a complete understanding of the differences between these models, we fit reading time data from all three conditions and parts of speech in Figure (4). From the coefficient estimates and their significance in the first and second rows, we observe significant facilitation—around 30 ms and 90 ms on average respectively in the "no" condition compared to the "wrong" condition, and in the "correct" condition compared to the others, in models fitted with text-based surprisals. This indicates that people are significantly faster in correct condition compared to other conditions and wrong condition significantly slows people down compared to not seeing any image at all. As before, we see that these effects, however, tend to shrink or disappear in models fitted with VLM surprisals(indicated with orange-pink on the diagram), while the impact of surprisal itself (along with other fixed predictors not shown in the figure) remains consistent across all models. **This strongly suggests that the notable difference in condition ID-related effects can only be explained by how the nature of surprisal changes when transitioning from text-based to multimodal models.** All this evidence also strongly indicates that Correct Image preview substantially affects comprehenders' expectations and that visual-language model surprisal captures a substantial part (though not all) of this effect.

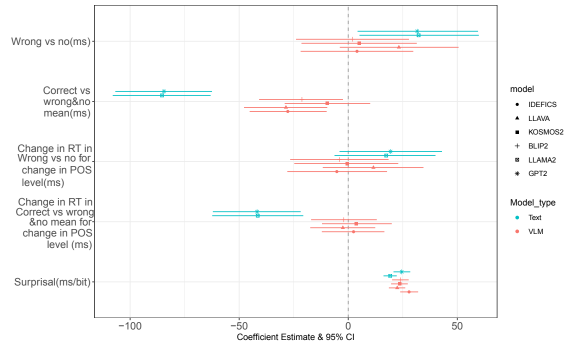

Figure 4: Coefficent Estimates and 95% CI of the fixed effects with theoretical interests. Note that the model had a Condition_ID*POS term, where Condition_ID was Helmert encoded making comparisons between wrong vs no and correct vs mean of wrong and no and POS was sum encoded with two levels, resulting in 2 interaction terms and 2 main effect terms for Condition_ID

6.2 Error Prediction

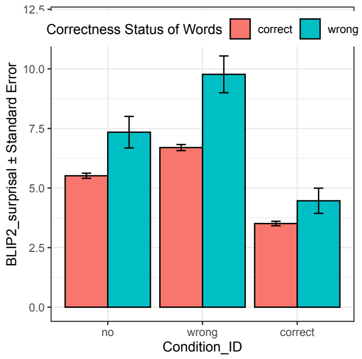

Figure 5: X axis indicates the conditions and correctness status of words(whether or not someone made a mistake in that word) and Y axis indicates mean and standard error of BLIP2 surprisal for words in a certain condition and correctness status

To investigate if the errors that people make have anything theoretically interesting to tell us, we first look into a univariate analysis showing the surprisal distribution across words in different conditions and correctness status. Consider the distribution of BLIP2 surprisal, which is a VLM, in figure (5). There is a very clear trend of high average contextual surprisal values for words that people got wrong. To prove this claim rigorously with a multivariate analysis, we fit a logistic regression model, so the goal is to predict the log-likelihood of making an error. Figure(6) shows the coefficient estimates and 95% confidence intervals of theoretically interesting predictors of this logistic regression model. From this figure, three things become evident - 1. From the first two rows, we see that the error occurrence likelihood does not vary much across different conditions, 2. From row

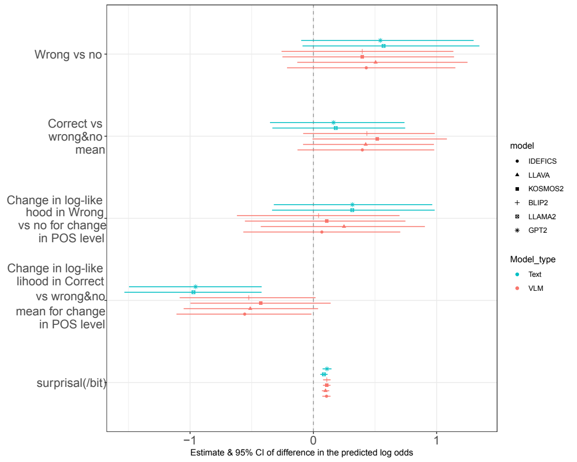

Figure 6: Estimate & 95% CI of difference in the predicted log odds of the fixed effects with theoretical interests. Note that the model had a Condition_ID*POS term, where the encoding of these terms is similar to before, resulting in 2 main effects of Condition_ID and 2 interaction terms, which is what we showed in the figure, along with surprisal.

4, we see that people are less likely to make errors for open parts of speech in the correct condition compared to other conditions (since the blue bars are on the negative side of the plot) and 3. From row 5, we see that the effect of surprisals is consistent across all models and increasing surprisal leads to more likelihood of error occurrence.

6.3 Can surprisal difference be explained as a function of groundedness?

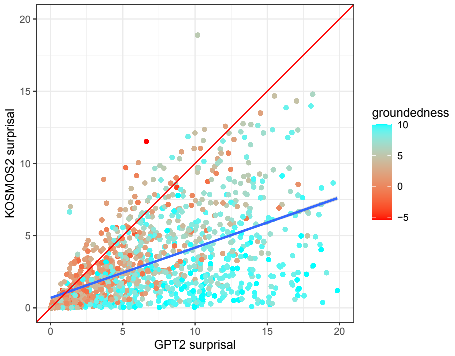

Figure 7: Every word token in every sentence in the dataset is indicated with a dot here. X coordinate of that dot indicates the GPT2 surprisal of that word given the previous words in that sentence and the Y coordinate of that dot indicates the KOSMOS2 surprisal of that word given the previous words and the image that sentence is describing, i.e, the KOSMOS2 surprisal in the correct condition. The color of the dot is determined by the groundedness rating of the word, noted as a scale to the right.

Consider the figure (7). We can notice that most dots below the dark blue line, the best-fitted linear relationship between GPT2 and KOSMOS2 surprisals, are light blue dots indicating highly grounded words. This motivation suggests that a lot of highly grounded words exhibit notably lower surprisal values in VLMs when contrasted with those derived solely from textual models. Intuitively speaking, ImageConditionedTextSurprisal minus TextSurprisal for a word roughly indicates the reduction of surprisal for the presence of the image. Hence, we expect that the more negative ImageConditionedTextSurprisal minus TextSurprisal is for a word, the more the effect of the image is on that word, hence the more grounded that word should be in the image. To formally analyze this nuance, in figure (8) we predicted the surprisal difference between two conditions from the same model using POS type, POS type and groundedness interaction, frequency and length as fixed effect predictors. In addition, we incorporated a random effect predictor that encompasses all fixed predictors, with the sentence type serving as the grouping variable. The significance of the groundedness effect on the surprisal difference for each type of POS is indicated such that “ns” means "not significant"; * means $p < 0.01$ and ** means $p < 0.001$.

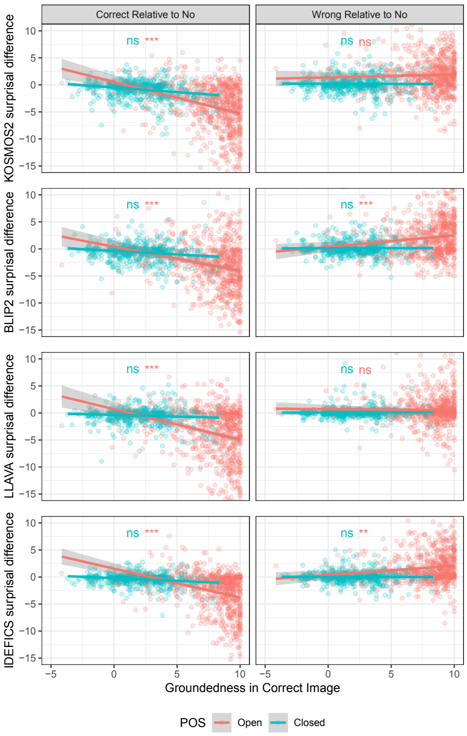

Figure 8: For each of the 4 VLMs we considered for this paper, the X axis indicates the groundedness value of a word and the Y axis indicates the difference between the surprisals of that word in correct condition and no condition (left panel) and wrong condition and no condition(right panel). The best linear fits for each type of POS(open/closed) are shown in the plots. The significance of groundedness contribution for each type of POS is also indicated in each plot.

Note that when comparing correct condition to no condition, we notice a consistent pattern of open class words' groundedness significantly contributing to the surprisal difference for all models. But we don't notice the same for closed class words, which makes sense given that they are mostly not strongly grounded in the image and hence the presence of an image doesn't give much extra information about them. **These findings highlight a strong correlation between human judgment of a word's degree of grounding in an image and the reduction in that word's surprisal for the presence of that image, as measured by recent VLMs.**

However, we notice a significant contribution of open class words' groundedness on surprisal difference between wrong and no conditions for BLIP2 and IDEFICS(but in the opposite direction of what we saw in the other comparison). At first, it might seem counter-intuitive but it just tells us that models like BLIP2 and IDEFICS struggle to ignore the image context in the wrong image condition, hence for the open class words in a sentence that would otherwise be grounded in the image in the 'Correct Image' context, they have significantly high surprisal due to those words' visual absence in the 'Wrong Image' context, resulting in the significance we observe in figure (8).

7 Perplexity and psychometric accuracy

In recent years, there has been an effort to study the increase of log-likelihood for including LLM surprisal estimate from models as a function of perplexity(Oh and Schuler, 2023). To investigate what traits in a VLM give them better predictive power for human RT, we ran a similar analysis with different-sized open-sourced versions of all the models we used in the work - two versions of all the VLMs except for KOSMOS-2 and a new VLM that improved upon Llava, Llava-Next. The baseline regression model was considered with all baseline predictors such as main effects of helmert encoded Condition_ID and sum encoded POS and interaction between them, frequency, length and full regression models additionally contained each LM surprisal predictor. Both the baseline and full regression models had the same random effects structure; a random intercept and slope for Condition_ID within each subject, sentence, word, and word token type was included. After fitting the regression models, we determined the increase in log-likelihood (ΔLL) for each model by subtracting the log-likelihood of the baseline model from that of the full model. Finally, the perplexity of each model type was calculated in our dataset of all items. Figure (9) shows the resultant plots.

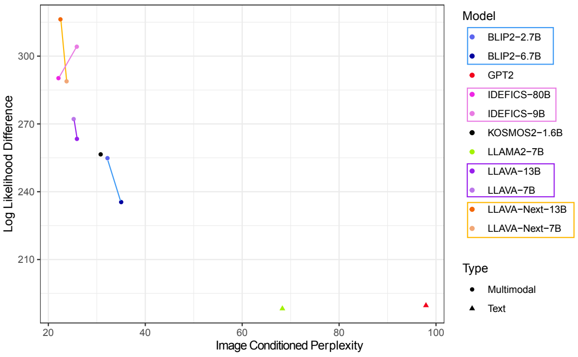

Figure 9: Increase in regression model log-likelihood fitted with data from all conditions for including each surprisal estimate as a function of **image-conditioned perplexity**, the different-sized versions of the same model are indicated with different shades of the same color and connected with a line for ease of interpretation.

Note that the increase of log-likelihood for adding surprisals from different-sized versions of the same model isn't very different, however different models can have very different predictive power regardless of the size, consider Llava and Llava-Next for example, both versions considered for these models have the same sizes(7B and 13B parameter) but Llava-Next has a lot more predictive power compared to Llava. This strongly indicates that training diet and objective are more important than the model size when it comes to psychometric predictive power. However, all the smaller-size versions except for Llava-Next are better than the bigger-size versions. Although this needs further exploration, the observations indicate that for each type of training objective and diet, there is possibly an optimal number of parameters that make the model most aligned with human expectations, and beyond that alignment decreases.

8 Conclusion

In this work, we have developed a novel experimental paradigm, Image-Conditioned Maze Reading, to study human linguistic expectations during real-time language comprehension when a visual context is involved. Our results demonstrate a substantial facilitatory effect of correct image context on language comprehension. This effect is evident not only for concrete nouns, adjectives, or verbs directly present in the image but also extends to words not explicitly grounded in the visual context. We extended psychometric benchmarking to visual language models and found that VLM surprisals capture most to all of the facilitator effect that occurs due to the presence of a relevant visual context. We discovered that as one goes from text based model surprisal to VLM surprisal, the effect of surprisal on reading time doesn't change much, but the huge Condition_ID related effects mostly disappear for VLM surprisal based models. So, the explanation is in how the nature of the surprisal changes. We also found a strong correlation between the human judgment of a word's degree of grounding in the image and the reduction of that word's surprisal for the presence of that image. We showed empirical support indicating that heightened contextual surprisal significantly contributes to errors in maze tasks. Finally, our findings reveal compelling evidence that the training objectives and diet of Vision-Language Models (VLMs) significantly impact their psychometric predictive power, more so than their size. However, this observation warrants further investigation.

9 Limitations

In this study, we used images and descriptions from the validation split of the COCO dataset. At that time, we were uncertain about the specifics of investigating Vision-Language Models (VLMs). Upon further examination down the line, we discovered that Llava and BLIP-2 had COCO in their pre-training data, indicating that these models may have encountered some of our items before. In future work, we plan to use images and descriptions from a dataset that has not been used for pre-training any of the models.

Another challenge we faced was the limited availability of different-sized versions of open-sourced VLMs for comprehensive analysis. There are typically only 2-3 versions available for each model. This limited our analysis compared to studies like (Oh and Schuler, 2023), which utilized many versions of Pythia models (Biderman et al., 2023) for interpretability analysis and understanding the development of knowledge in autoregressive transformers. The scarcity of multiple versions of open-sourced VLMs hindered our ability to perform a similarly comprehensive analysis.

References

Paul D. Allopenna, James S. Magnuson, and Michael K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38:419–439.

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

Richard Antonello, Aditya Vaidya, and Alexander Huth. 2023. Scaling laws for language encoding models in fmri. In *Advances in Neural Information Processing Systems*, volume 36, pages 21895–21907. Curran Associates, Inc.

S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. ... Hallahan, and O. Van Der Wal. 2023. A suite for analyzing large language models across training and scaling. *In International Conference on Machine Learning (pp. 2397-2430). PMLR.*

V. Boyce and R. Levy. 2023. A-maze of natural stories: Comprehension and surprisal in the maze task. *Glossa Psycholinguistics, 2(1).*

Veronica Boyce, Richard Futrell, and Roger Levy. 2020. Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:1–13.

M. Brysbaert and B. New. 2009. Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods, 41, 977-990.*

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134.

Delphine Dahan, James S Magnuson, and Michael K Tanenhaus. 2001. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, 42(4):317–367.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. 20:641–655.

Kenneth I Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, 41(1):163–171.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. pages 61–69, Montreal, Quebec.

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. pages 159–166, Pittsburgh, Pennsylvania.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. 2022. A hierarchy of linguistic predictions during natural language comprehension. 119(32):e2201968119.

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.

Marta Kutas and Kara D Federmeier. 2011. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647.

Marta Kutas and Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, ... Lozhkov, A., and V. Sanh. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems, 36.*

Roger Levy. 2008a. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy. 2008b. A noisy-channel model of rational human sentence comprehension under uncertain input. pages 234–243, Waikiki, Honolulu.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597.*

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

H. Liu, C. Li, Y. Li, and Y. J. Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744.*

William Marslen-Wilson. 1975. Sentence perception as an interactive parallel process. *Science*, 189(4198):226–228.

B. D. Oh and W. Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. *Conference on Empirical Methods in Natural Language Processing.*

Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824.*

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. 124(3):372–422.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger P. Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. 121(10):e2307876121.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition, 128:302–319.*

Jakub M Szewczyk and Kara D Federmeier. 2022. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, 123:104311.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. ... Babaei, and T. Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Naoko Witzel, Jeffrey Witzel, and Kenneth Forster. 2012. Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, 41:105–128.

J. Zehr and F. Schwarz. 2018. Penncontroller for internet based experiments (ibex). *https://doi.org/10.17605/OSF.IO/MD832*.