**Expectation-based comprehension of linguistic input: facilitation from visual context**

**Background:** Context fundamentally shapes real-time human language processing, creating linguistic expectations that drive efficient processing and accurate disambiguation (Kuperberg and Jaeger, 2016). In naturalistic language understanding, the visual scene often provides crucial context (Ferreira et al., 2013; Huettig et al., 2011). We know that visual context guides spoken word recognition (Allopenna et al., 1998), syntactic disambiguation (Tanenhaus et al., 1995), and prediction (Altmann and Kamide, 1999), but much about how visual context shapes language understanding remains unknown. Here we investigate how visual context influences linguistic expectations using an experiment in which participants preview images for 5 seconds and then read English-language image descriptions in the Maze task (ideal here due to its high sensitivity and low spillover) (Forster et al., 2009; Boyce et al., 2020), in one of three conditions (Figure 1): **No Image** preview; **Correct Image** preview; or **Wrong Image** preview. We hypothesize that image preview will rapidly shape linguistic expectations, facilitating comprehension in the Correct Image condition relative to the other two conditions. We distinguish between a **lexical-grounding** hypothesis, in which this effect holds only for words in open-class parts of speech referring directly to objects, properties, or events depicted in the image, and a **comprehensive-grounding** hypothesis, in which this effect holds for all words regardless of part of speech. We independently norm each word's degree of grounding in the correct image, and estimate word-by-word image description surprisal using two autoregressive large language models (LLMs): GPT–2 (Radford et al., 2019), which uses only linguistic context, versus BLIP–2 (Li et al., 2023), a recent multimodal LLM that additionally conditions on an image and is trained to generate image descriptions.

**Methods:** We chose 108 image–description pairs from the COCO dataset (add citation), randomly reshuffling them to create Wrong Image pairs. Each participant (n=69 from Prolific, based on pilot data and power analysis) participated in 36 trials, 12 in each of the three conditions, with items for each trial randomly sampled without replacement. We analyzed RTs with Bayesian mixed linear models, with word length and frequency as control predictors.

**Results:** LLM comparison shows that correct-image conditioning substantially reduces surprisal, especially for the words most grounded in the image (Figure 2). This groundedness–surprisal relationship is stronger for open-class words than for closed-class words. Maze RTs are faster in the Correct Image condition than in No Image and Wrong Image conditions (Figure 3), not only for open-class words (both $p < 0.001$) but also for closed-class (both $p < 0.001$); facilitation is greater for open-class than for closed-class words (both $p < 0.001$). Only words well-grounded in the image show significant facilitation, and the grounding–facilitation relationship is the same for open- and closed-class words (Figure 4). GPT–2 surprisal additionally predicts RTs, but its effect size is much smaller with correct-image than with wrong-image preview (Figure 5). When BLIP–2 surprisal is used instead, surprisal effect sizes are the same in the two conditions, and the Correct Image groundedness effect size is substantially reduced. These results suggest that Correct Image preview substantially affects comprehenders' expectations, and that BLIP–2 surprisal captures a substantial part (though not all) of this effect.

**Discussion:** Our results support the **comprehensive-grounding** hypothesis: we see Correct-Image facilitation for words of all parts of speech (Figure 3), with the amount of facilitation modulated by the word's degree of grounding in the correct image (Figure 4). Our results also suggest that these effects may be largely mediated by the effect of image context on surprisal. Our work also offers new possibilities for how multimodal large language models may be used in psycholinguistic research to investigate how visual context affects language processing.

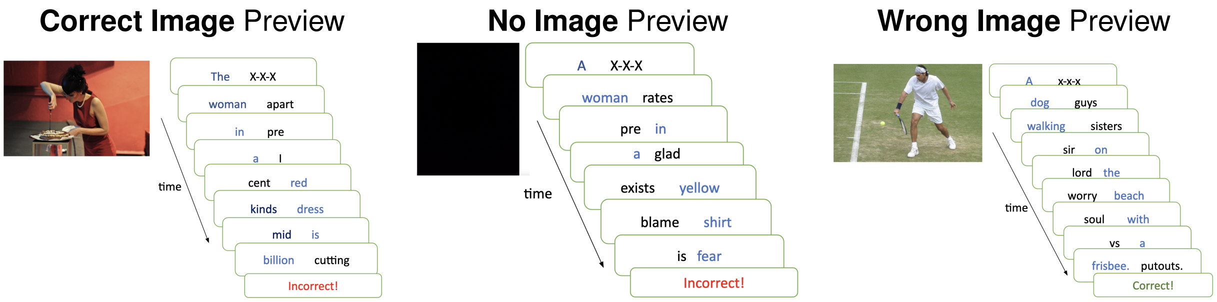
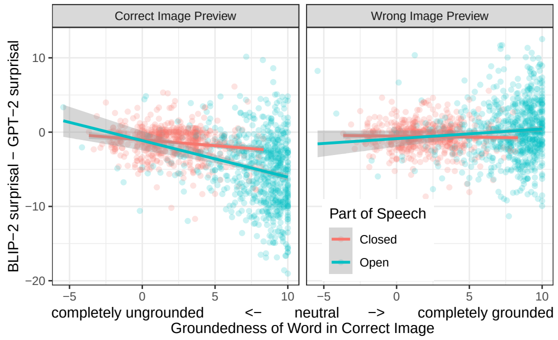


Figure 1: Schematic of image-description A-maze reading in each of the three experimental conditions. Participants first briefly view an image and then read a description by successively choosing the word fitting the preceding linguistic context and rejecting a foil word (example selections marked in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



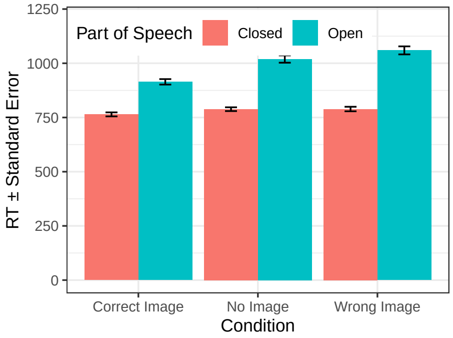
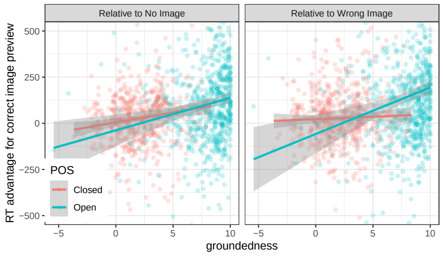


Figure 2: Image preview surprisal effect (difference between BLIP-2 & GPT-2) and its relationship with word grounding

Figure 3: Mean word RT, as a function of image-preview condition and part of speech



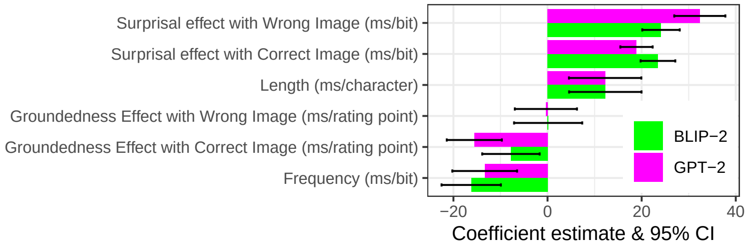


Figure 4: Correct image preview advantage, by word groundedness and part of speech

Figure 5: Mixed linear regression coefficients analyzing RTs using GPT–2 surprisal vs BLIP–2 surprisal

**References** • Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (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. • Altmann, G. T. and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3):247–264. • Boyce, V., Futrell, R., and Levy, R. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, 111:1–13. • Ferreira, F., Foucart, A., and Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, 69(3):165–182. • Forster, K. I., Guerrera, C., and Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, 41(1):163–171. • Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2):151–171. • Kuperberg, G. R. and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1):32–59. • Li, J., Li, D., Savarese, S., and Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597.* • Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. • Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.