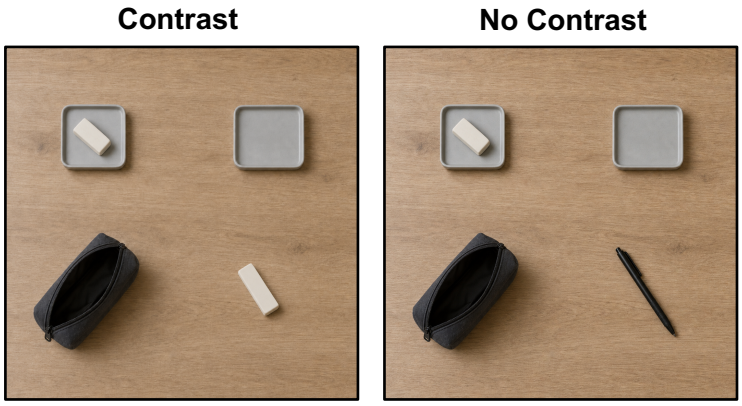
Syntactic Disambiguation from Visual Context in Vision-Language Models

Introduction. Vision-Language transformer Models (VLMs) use the visual information in an input image to inform their linguistic predictions [1-3]. However, it is unknown whether visual context can modulate a VLM's incremental syntactic parsing (and subsequent interpretation) of a structurally ambiguous sentence. In humans, visual inputs can rapidly support such disambiguation, as demonstrated most famously by Tanenhaus et al. [4]. Here, we adapted their design to VLMs: do these models show human-like use of visual context for syntactic disambiguation during incremental processing?

Methods. We generated 60 naturalistic, minimal pairs of images with one object positioned in each of the four image quadrants (**Fig. 1**). Each pair included: (i) a "contrast" image, containing a target object on a "platform" (e.g., an eraser on a tray), a copy of that target object with no platform (an eraser), a copy of the platform with nothing on it (empty tray), and a container object (a bag); and (ii) a "no contrast" image, identical to the contrast image except that a distractor object (e.g., a pen) replaced the target object that had no platform. Images were paired with two sentence stems: Reduced relative clauses and Unreduced relative clauses, e.g., "*put the eraser (that's) on the tray*". Using 8 different templates of these two conditions, we obtained a total of 8 × 60 = 480 text-image stimuli per condition. In the critical, Reduced condition, the prepositional phrase ("*on the tray*") is ambiguous between a modifier of "*eraser*" and its intended destination; without context, humans interpret it as a destination because the verb ("*put*") requires one. However, the "contrast" image biases humans towards a modifier reading that specifies the referent (the eraser that sits on the tray). The control (Unreduced) condition has an unambiguous modifier reading.

We analyzed two transformer-based VLMs: LLaVA [1] and InstructBLIP [2]. They were fed each text-image stimulus, and we measured the surprisal of the preposition "*in*" as the next word: if the stem's PP ("*on the tray*") is read as a destination, a following "*in*" would be surprising; however, "*in*" would be expected if the stem's PP is read as a modifier, because in that case the main verb ("*put*") would be awaiting a destination. We thus predicted that for Reduced (critical) sentences, surprisal would be higher for "no contrast" images (bias towards a destination reading) compared to "contrast" images (bias towards a modifier reading); we predicted no difference for the unambiguous, Unreduced sentences. These effects should further yield an Image Type × Clause Type interaction (**Eq. 1**). For the ambiguous, Reduced condition, we also measured the surprisal of the sentence ending directly after "tray" (surprisal of '.' or '?', depending on the stem's template version), predicting that it would behave oppositely to "*in*": higher for "contrast" images (because a modifier reading of the stem's PP leaves the verb with no destination) than for "no contrast" images. This pattern would yield an Image Type × Token Type interaction (**Eq. 2)**.

Results and Discussion. As predicted, surprisal for the continuing preposition "*in*" was higher for "no contrast" images than for "contrast" images within Reduced sentences (LLaVA: b=0.26, SE=0.044, $t_{(59)}$=6.04, p<.0001; InstructBLIP: b=0.090, SE=0.032, $t_{(93.9)}$=2.81, p=.003), but not within Unreduced (control) sentences (LLaVA: b=-0.11, SE=0.045, $t_{(59)}$=-2.41, p=0.99; InstructBLIP: b=0.055, SE=0.032, $t_{(93.9)}$=1.73, p=.044; $\alpha = 0.05/2$). Thus, both VLMs show human-like use of visual context for syntactic disambiguation (**Fig. 1**). However, the predicted Image Type × Clause Type interaction was only significant in LLaVA (b=0.37, SE=0.037, $t_{(59)}$=10.12, p<.0001) (InstructBLIP: b=0.0035, SE=0.0029, $t_{(1725)}$=1.19, p=0.12). As predicted, for Reduced sentences, we found that image type had different effects on the surprisal of "*in*" vs. the surprisal of the end-of-sentence punctuation (LLaVA: b=0.32, SE=0.062, $t_{(58.99)}$=5.086, p<.0001; InstructBLIP: b=0.0091, SE=0.0037, $t_{(1850)}$=2.49, p=0.0064). Yet, contrary to our hypothesis, the surprisal of punctuation did not significantly differ by image type (LLaVA: b=-0.054, SE=0.032, $t_{(59)}$=-1.67, p=0.05; InstructBLIP: b=-0.0011, SE=0.026, $t_{(1850)}$=-0.044, p=0.26). Therefore, whereas a modifier reading is more likely in "contrast" than "no contrast" images, a destination reading is no more likely in "no contrast" than "contrast" images. These data suggest that VLMs partially align with humans in their use of visual context to inform syntactic processing.



Text Stimuli (Clause Type):
Reduced: Put the eraser on the tray [...in]
Unreduced: Put the eraser that's on the tray [...in]

Figure 1: An example pair of images and text stimuli. The two image conditions and clause types were crossed in a 2 × 2 design.

Statistical Models:

(1) Surprisal ~ Image Type × Clause Type + Target Quadrant +
(Image Type × Clause Type | Sentence Type) +
(Image Type × Clause Type | Item)

(2) Surprisal ~ Image Type × Token Type + Target Quadrant +
(1 | Sentence Type) + (1 | Item)

Equations 1-2 Linear Mixed-Effects Models. Image Type: "contrast" vs. "no contrast"; Clause Type: Reduced vs. Unreduced; Target Quadrant: the location of the target object (without the platform) in the image; Sentence Type: 8 template versions (all paraphrases of the two clause types); Item: 60 minimal pairs of images; Token Type: "in" vs. punctuation. The random effects show the maximal model structure that converged. When models did not converge, we removed the random effects with the lowest variance until convergence.

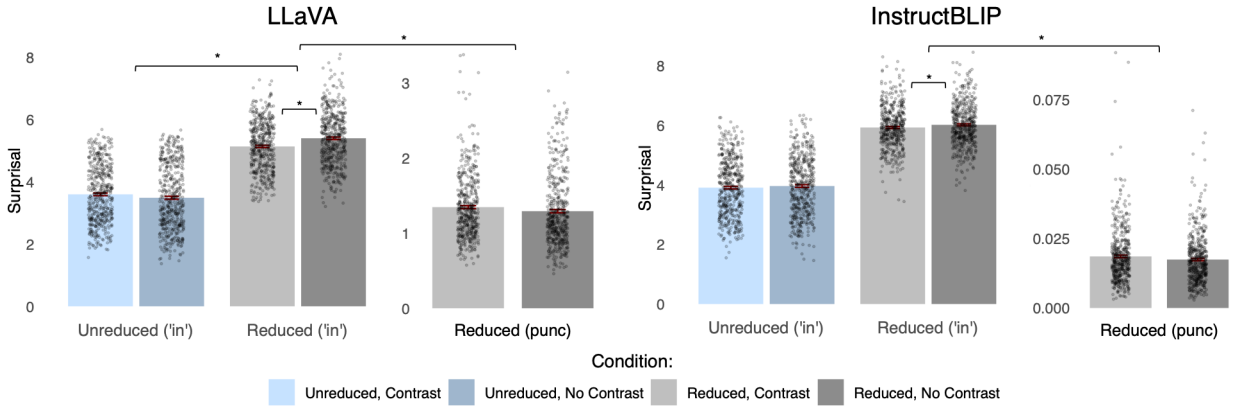


Figure 2. Bar plots showing average surprisal for the critical preposition ("*in*") and end-of-sentence punctuation within levels of Image Type ("contrast" vs. "no contrast") and Clause Type (Reduced vs. Unreduced). Dots are individual stimuli. Error bars show ± 1 standard error of the mean.

References. [1] Liu, Li, *et al.* (2023), *NeurIPS*; [2] Dai *et al.* (2023), *NeurIPS*; [3] Berzak *et al.* (2015), *EMNLP*; [4] Tanenhaus *et al.* (1995), *Science*.