Multimodal Pragmatic Inference in Vision-Language Transformers

Thomas McGee[1], Meng Du[1], Megan Jacob[3], Idan Blank[1,2]

[1] Department of Psychology, University of California, Los Angeles, Los Angeles, USA
[2] Department of Linguistics University of California, Los Angeles, Los Angeles, USA
[3] Department of Computer Science, University of California, Los Angeles, Los Angeles, USA

Abstract

Contemporary transformer models have achieved human-like performance on many text-based tasks. However, real-world communication requires the integration of linguistic representations with contextual information from other domains (e.g., visual, social, etc.). Here, we study such information integration in three multimodal transformer models (LLaVA, InstructBLIP, and GPT-4o). We test these models' pragmatic inference capabilities regarding referring expressions: when an array of objects contains two exemplars from the same category that differ in size, referring to one of them felicitously requires a size adjective (e.g., "the big hammer"); However, that adjective is unnecessary if only one exemplar from the category is present. We evaluate such inferences (1) when models process text-image inputs, where we extract multimodal surprisal for critical words and compare them across infelicitous vs. felicitous expressions; and (2) when models generate descriptions of input images given text prompts. We find evidence for pragmatic integration of visual and linguistic context in all models. However, these inferences remain sensitive to the in-context statistics of visual inputs—an important difference from pragmatic inference in humans.

Keywords: Pragmatics; language comprehension; language production; predictive processing; artificial intelligence

Introduction

A fundamental criterion for evaluating artificial intelligence (AI) systems is the degree to which they align with human communicative norms (Linzen, 2020; Turing, 1950; Wang et al., 2019). Large language models (LLMs) have reached near-human levels of performance on tasks that require literal interpretation of language (Chang & Bergen, 2024; Gauthier et al., 2020; Hu et al., 2020; Piantadosi, 2023), including syntactic parsing (Contreras Kallens et al., 2023; Linzen & Baroni, 2021; Wilcox et al., 2021) and semantic processing (Chronis & Erk, 2020; Petersen & Potts, 2023). Beyond such "formal" competence (Mahowald, Ivanova et al., 2024), human-like real-world communication crucially relies on the non-literal (pragmatic) interpretation during both comprehension (Degen & Tanenhaus, 2015; Goodman & Frank, 2016; Papafragou & Musolino, 2003; Ryskin et al., 2019; Sedivy et al., 1999) and production (Do et al., 2020; Gatt et al., 2017; Koolen et al., 2013; Rubio-Fernandez et al., 2025; Sedivy, 2003, 2004). Example cases include figurative language use (e.g., Glucksberg, 2003; Holyoak & Stamenković, 2018), irony, hyperbole, and vagueness (Goodman & Frank, 2016). Language models appear capable of such complex, non-literal inferences from text (Hu et al., 2023; Ichien, Stamenković, & Holyoak, 2024; Ruis et al., 2022; Stowe, Utama, & Gurevych, 2022; Tong, Shutova, & Lewis, 2021).

However, it remains unknown whether artificial neural networks are capable of pragmatic inferences that require multimodal integration of language with non-linguistic, contextual information (Grodner & Sedivy, 2011; Noveck & Reboul, 2008; Sedivy et al., 1999; Tanenhaus et al., 1995). Contextual cues, such as visual information, support the mutual inferences that interlocutors make about each other's intentions and knowledge (e.g., Grodner & Sedivy, 2011; Ryskin et al., 2019). Here, we focus on one such inference: reasoning about how people refer to the objects around them.

One pragmatic principle for such reasoning is the expectation that one's conversation partner would adhere to the Gricean maxim of quantity (Grice, 1975; Sedivy et al., 1999): be as informative as necessary (including when referring to an object), but not more than required. For instance, imagine that the visual context of a conversation includes a table with a big plate, a small plate, a big box, and a small ball. A partner who wants the former plate should not say "*give me the plate*"—that utterance is under-informative (which plate?). Instead, they should use disambiguating information, such as a size adjective: "*give me the big plate*".

Indeed, humans are more likely to produce such size adjectives when an array of objects includes two exemplars from the same category that differ in size (a "size contrast") than when there is only one exemplar per category (e.g., Brown-Schmidt & Konopka, 2011; Brown-Schmidt & Tanenhaus, 2006; Sedivy, 2003). Conversely, human comprehenders integrate such visual context as soon as it becomes available to support incremental comprehension. Such integration has been demonstrated by tracking listeners' eye movements as they inspect an object display and follow spoken instructions, in a method called the "visual world paradigm" (Cooper, 1974; Sedivy et al., 1999; Tanenhaus et al., 1995). Upon observing the visual array {big plate, small plate, big box, small ball} and hearing "*Click on the big…*", people are more likely to make an anticipatory saccade to the plate (a member of the size contrast) than to the box (the other big object) before hearing the final noun (Grodner & Sedivy, 2011; Ryskin et al., 2019; Sedivy, 2003; Sedivy et al., 1999).

Therefore, here we asked: can we find human-like patterns of vision-language integration to support pragmatic reasoning in multimodal transformer LLMs that process both text and images? Whereas previous work has tested pragmatic inferences limited to contextual factors that can be

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

learned from text (Cho & Kim, 2024; Cong, 2024; Lin et al., 2024), we are not aware of any studies testing multimodal models on the much more challenging task of cross-domain information integration. Our study thus moves beyond text-only benchmarks towards a fuller evaluation of LLM's communicative competence.

There are reasons to suspect that multimodal transformers might have emergent pragmatic-like abilities. First, contemporary models dedicate substantial architectural resources to learning multimodal representations (i.e., numerous multimodal attention and multi-layer perceptron layers; Dai et al., 2023; Li et al., 2023; Liu, Li et al., 2023; Radford et al., 2019). Given this extensive parameterization for vision-language integration, these models may be capable of learning nuanced multimodal representations. Second, the training objectives of multimodal transformers, such as text-image instruction-following (Liu, Li et al., 2023) and image-grounded text generation (Li et al., 2023) may constrain them towards representing and processing language in alignment with human communicative norms.

We conducted two experiments, to test pragmatic inferences during both "comprehension" (input processing) and production. The first experiment presents multimodal transformers with images of the kind typically used in the visual-world paradigm, along with instructions regarding the image. We convert the measure of predictive processing from anticipatory looking in humans to the (un)predictability that the model assigns to critical words, estimated via surprisal (Hale, 2001; Levy, 2008). We compare surprisal for the same instruction between minimal pairs of images differing in a single object (e.g., with vs. without a size contrast) or, given a single image, between minimal pairs of instructions differing in a single word. We hypothesized that surprisal would be higher for instructions that are infelicitous (i.e., over- and under-informative) than for felicitous instructions.

The second experiment uses similar stimuli to assess the pragmatic felicity of model-generated descriptions via an open-ended generation task. We hypothesized that multimodal transformers would use size adjectives to modify a target object that they are referring to more frequently when that object is part of a contrast set than when it is not.

Experiment 1: Processing Input

Methods

Model Description. We analyzed two multimodal models: LLaVA (Liu, Li et al., 2023) and InstructBLIP (Dai et al., 2023). Both are transformer-based vision-language models, combining an image encoder, a text-only large language model, and a cross-modal module that integrates image and text. LLaVA uses a simple linear layer to map image features extracted by the image-encoder into the text embedding space used by the language model. InstructBLIP first uses a "Query transformer" to extract task-relevant image features from the output of the image encoder, which are then linearly projected onto the text embedding space.

Analysis 1: Felicitous vs. Over-informative. We used text-image pairs as stimuli. Each image contained four objects, one in each quadrant, randomly selected from Ryskin et al. (2019). Two objects were big and two were small. Images were 1100×1100 pixels in size; big objects were 500×500 pixels in size, and small objects—150×150 pixels (objects did not fully occupy these pixels; see Figure 1). Each object was centered within its respective quadrant.

Images were created in 60 sets. Each set consisted of 3 image types with a shared set of objects. These versions were paired with text inputs, giving rise to 4 conditions (Figure 1):

1. **Size contrast, felicitous**: the image had a size contrast (e.g., big and small hammers), one big object (e.g., cork), and one small object (e.g., lobster). The instruction referred to an object from the contrast set, using a size adjective ("*In the image, point to the big hammer"*).
2. **Size contrast, over-informative**: the same image as in (1), but the instruction referred to an object that was not part of the contrast set, using a size adjective ("*In the image, point to the big cork"*).
3. **Different size contrast, over-informative**: the image included the same object categories, but now a different category was the basis of a size contrast (e.g., big and small cork; big hammer; small lobster). The instruction was the same as in (1), thus using a size adjective to refer to an object that was now not part of a contrast set.
4. **No size contrast, over-informative**: The image had objects from 4 distinct categories, with no size contrast (e.g., big hammer, big cork, small lobster, small boots). The instruction was the same as in (1).

The object referred to in (1), (3), and (4) appeared in the same quadrant across all versions within a set, but its quadrant was counterbalanced across sets. The quadrants of the remaining objects were randomized.

We fed each text-image pair to LLaVA and InstructBLIP, and quantified the unpredictability of the last, critical word (i.e., the object noun) given the multimodal context.[1] To this end, we first computed the next-token probability distribution at the size adjective by applying the softmax operation over extracted logit scores; then, we took the negative logarithm of the probability for critical token *x*:

$$surprisal = -log_2\, p(x \mid context)$$

In cases where the critical noun was a multi-token word or an open compound noun, we computed probabilities incrementally for each token constituting the noun, incorporating all preceding tokens as context. Then, we converted token probabilities to surprisal values and summed them to obtain surprisal for the entire word (or compound).

[1] For InstructBLIP, we excluded the text prompt from the Q-Former input (which uses bidirectional attention), to prevent embeddings passed to the decoder from encoding information about future tokens.

We compared surprisal for the critical noun between Condition 1 (felicitous) and each of the 3 over-informative conditions, using Bonferroni-corrected, one-tailed, paired samples *t*-tests. Conditions 1 and 2 differ in the critical word but have an identical image; Conditions 1, 3, and 4 use minimally different images but have identical text. Higher surprisal for the infelicitous conditions (2-4) than the felicitous condition (1) would resemble pragmatic inference.

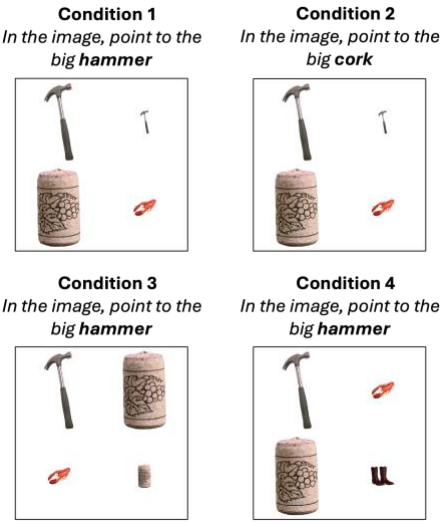

Figure 1: The conditions used in Surprisal Analysis 1.

Analysis 2: Felicitous vs. Under-informative, and Control for In-Context Object Frequency. We used identical images to those in Analysis 1 but modified the text so that we could compare felicitous referring expressions to *under*- (rather than *over*-) informative ones. To this end, we removed the size adjectives from all sentences, which reversed the pragmatic felicity of each condition. It made Condition 1 under-informative (i.e., an unmodified noun has two viable visual referents, i.e., "*In the image, point to the hammer.*" in the presence of two hammers), and Conditions 2-4 felicitous (i.e., an unmodified critical noun refers to a single visual referent). All other aspects of the experiment were identical to those in Surprisal Analysis 1. Higher surprisal for the infelicitous condition (1) than the felicitous conditions (2-4) would resemble pragmatic inference.

Unlike in Analysis 1, each sentence ended with a period that was included as a token in the multi-token surprisal calculation. The period ensured that Condition 1 was indeed infelicitous (otherwise, a disambiguating size adjective could follow the critical noun, rendering the sentence felicitous, e.g. "*In the image, point to the hammer that is big"*). Similarly, it ensured the felicity of Conditions 2-4 (otherwise, an over-informative size adjective could follow the critical noun).

This analysis also serves as a control for a confound in Analysis 1, where pragmatic inference makes the same predictions as expecting that the text would simply refer to the most frequent object in the image: in the felicitous condition (1), the noun refers to the most frequent object, so surprisal should be low; in the infelicitous conditions, it refers to an object other than the most frequent one (2, 3), or all objects have the same frequency (4), so surprisal should be high. In contrast, in Analysis 2, pragmatic inference and in-context object frequency make opposite predictions.

Analysis 3: Control for In-Context Object Frequency. We also controlled for the object-frequency account described above using two new conditions (Figure 2) that tested the same image with two different texts. The image showed a target object in a size contrast and two equally sized, identical copies of a different object (e.g., big hammer, small hammer, and two big corks). In Condition 1, the text was felicitous (referring to "*the big hammer.*"); in Condition 2 it was under-informative (referring to "*the big cork.*").

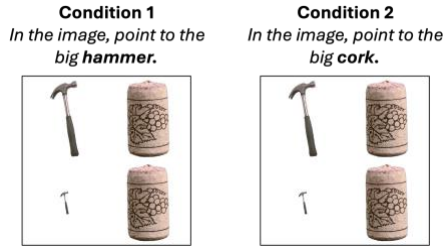

Figure 2: The conditions used in Surprisal Analysis 3.

As in Analysis 2, each sentence ended with a period, and surprisal values were summed for the noun and the period. Periods ensured that Condition 2 was infelicitous (otherwise, information could follow the critical noun to disambiguate which cork is the referent, increasing pragmatic felicity).

Higher surprisal for the under-informative condition (2) than the felicitous condition (1) would resemble pragmatic inference. A different pattern is predicted by the in-context object frequency account: the image contains two hammers and two corks, so there is no difference in frequency. In fact, there are two big corks and only one big hammer, so "*big cork*" refers to a more frequent object than "*big hammer*". Therefore, surprisal should either be similar between the two conditions, or lower in Condition 2.

Analysis 4: Felicitous vs. Over-informative, Surprisal for the Size Adjective. We compared Conditions 1 (contrast set) and 4 (no contrast set) from Analysis 1 but extracted surprisal for the size adjective (from next-token probabilities at "*the*") instead of the noun. A size adjective would only be felicitous when a contrast set is present, as in Condition 1, so higher surprisal for the over-informative Condition 4 would resemble pragmatic inference.

Analysis 5: Pragmatic inference with no size adjectives. In humans, the expectation for size adjectives in the context of contrast sets is a pragmatic inference. In the models we test, it might instead reflect a lexical semantic effect: the meaning of size adjectives may indicate a comparison class (Sedivy, 2003). Thus, we tested whether pragmatic-like effects could

be elicited without size adjectives. Images contained two equally sized, identical copies of an object (e.g., two hammers). In Condition 1 (felicitous), the text used a plural noun ("*In the image, look at the hammers.*"); in Condition 2 (under-informative), it used a plural noun ("*look at the hammer.*"). We used the verb "look" instead of "point" in this analysis, because a command to look at multiple objects seemed more pragmatically appropriate than a command to point at multiple objects. Higher surprisal for the singular noun (including the period) would resemble pragmatic inference.

Results

Analysis 1: Felicitous vs. Over-informative. As predicted (Figure 3), compared to felicitous Condition 1, LLaVA exhibited higher surprisal for the critical noun in the infelicitous Condition 2 (mean difference (M)=2.81 bits, $t_{(59)}$=3.06, p=0.0017), Condition 3 (M=2.66, $t_{(59)}$=6.43, $p<10^{-7}$), and Condition 4 (M=2.77, $t_{(59)}$=6.11, $p<10^{-7}$). Similarly, in InstructBLIP, compared to Condition 1, surprisal was higher in all other conditions (2 vs. 1: M=3.61, $t_{(59)}$=4.39, $p<10^{-4}$; 3 vs. 1: M=2.82, $t_{(59)}$=6.39, $p<10^{-7}$; 4 vs. 1: M=3.28, $t_{(59)}$= 6.60, $p<10^{-8}$). These data are consistent with pragmatic inference.

Analysis 2: Felicitous vs. Under-informative, and Control for In-Context Object Frequency. Our results showed no evidence for pragmatic reasoning in either model (Figure 3). Descriptively, surprisal patterned in the opposite way to our prediction: it was lower for the infelicitous (under-informative) condition than the felicitous conditions. To test the significance of this opposite pattern, we conducted two-tailed t-tests. In LLaVA, compared to infelicitous Condition 1, surprisal at the critical noun was significantly higher in felicitous Condition 3 (M=2.28 bits, $t_{(59)}$=5.72, $p<10^{-6}$) and Condition 4 (M=2.50, $t_{(59)}$=5.61, $p<10^{-6}$), but not in Condition 2 following Bonferroni correction (M=1.91, $t_{(59)}$=2.40, p=0.02). Similarly, in InstructBLIP, compared to Condition 1, surprisal was significantly higher in Condition 2 (M=3.63, $t_{(59)}$= 4.55, $p<10^{-4}$), Condition 3 (M=2.71, $t_{(59)}$= 6.26, $p<10^{-7}$), and Condition 4 (M=3.76, $t_{(59)}$= 7.06, $p<10^{-8}$). These results are inconsistent with pragmatic inference, and instead support an expectation that the text will refer to the most frequent object in the image.

Analysis 3: Control for In-Context Object Frequency. As predicted, surprisal for the critical noun was lower when it referred to a referent in a contrast set (e.g., a large hammer in the presence of a small hammer) than when it ambiguously referred to one of two equally sized, identical objects (e.g., large corks). This pattern held in both LLaVA (M=2.23 bits, $t_{(59)}$=2.40, p=0.0098) and InstructBLIP (M=2.77, $t_{(59)}$= 3.00, p=0.0022), and it is consistent with pragmatic inference even when controlling for in-context object frequency.

Analysis 4: Felicitous vs. Over-informative, Surprisal for the Size Adjective. As predicted, surprisal for the size adjective was higher in the infelicitous, over-informative Condition 4 (no contrast set) than the felicitous Condition 1 (contrast set), in both LLaVA (M=0.60 bits, $t_{(59)}$=5.45, $p<10^{-6}$) and InstructBLIP (M=0.29, $t_{(59)}$ = 2.81, p = 0.0034). These data are consistent with pragmatic inference.

Analysis 5: Pragmatic inference with no size adjectives. In LLaVA, as predicted, surprisal for the critical noun was higher in the under-informative condition than the felicitous condition (M=0.47 bits, $t_{(59)}$=2.08, p=0.021). These data are consistent with pragmatic inference. However, in InstructBLIP, this difference was non-significant (M=-0.12, $t_{(59)}$= -0.54, p=0.71).

Comparison of Effect Sizes between Analyses 1 and 2. We found evidence consistent with pragmatic inference in Analysis 1, but evidence in the opposite direction in Analysis 2. We therefore compared effect sizes across these analyses. If pragmatic effects were present in Analysis 2 but were overpowered by the effects of in-context (visual) object frequency, differences between conditions should be higher in Analysis 1 than in Analysis 2. In LLaVA, this was the case for Conditions 1 vs. 2 (mean difference of differences=0.90 bits, $t_{(59)}$=2.41, p=0.010) and 1 vs. 3 (mean=0.38, $t_{(59)}$=2.22, p=0.015), but not for Conditions 1 vs. 4 (mean= 0.27, $t_{(59)}$ = 1.73, p = 0.045). For InstructBLIP, effect sizes for between-condition comparisons did not significantly differ between Analyses 1 and 2 (all ps > 0.29).

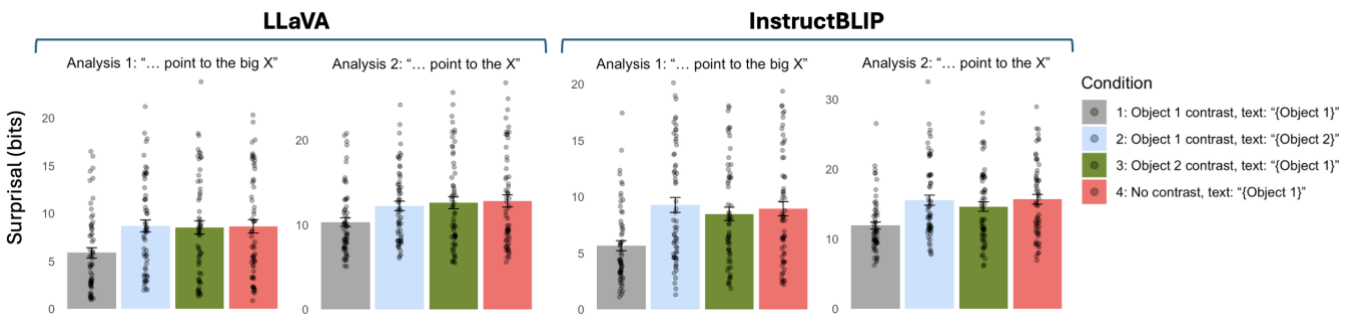

Figure 3: Surprisal values for the critical noun in each condition for Analyses 1 and 2. Bars show means, dots show data for single text-image pairs, error bars show standard errors. Data are shown for LLaVA (left) and InstructBLIP (right).

Experiment 2: Producing Text

Methods

Model Description. We analyzed three multimodal models: GPT-4o (OpenAI, 2024), LLaVA (Liu, Li et al., 2023), and InstructBLIP (Dai et al., 2023). GPT-4o is a single model that processes and generates image, audio, and text, but specific details about its architecture and training have not been released. For details on InstructBLIP and LLaVA, refer to Experiment 1.

Materials and procedure. We used images from Conditions 1 (size contrast) and 4 (no size contrast) in Experiment 1. Images were fed to GPT-4o along with the following prompt: *Given this image, your task is to describe one of the objects to a listener. Describe the object in the {quadrant} quadrant, so that a listener would know which object to look at. Please respond by briefly completing this sentence: 'Look at the '.* The position of the target object was denoted by *{quadrant}*: top left, top right, bottom left, or bottom right. We used the default values for the top_p and sampling temperature (1.0 for both). The top_p hyperparameter determines the probability mass from which the next token can be generated, with 1 corresponding to the entire distribution (OpenAI, 2023). Temperature controls the randomness of next-token sampling (OpenAI, 2023).

Because LLaVA and InstructBLIP made frequent errors in naming the correct target in their responses to the prompt above, we used a slightly different prompt for them, which stated the target object's name. It started with: *Your task is to describe the object in the {quadrant} quadrant ({object}), so that a listener* [same continuation]. The prompt did not include a size adjective. We used deterministic sampling, in which the single most probable token is produced at each inference step.

We annotated models' responses for the use of size adjectives. We predicted that size adjectives would be produced more frequently when the target object was part of a contrast set, which would resemble pragmatic reasoning. To ensure that such responses do not instead reflect a general tendency to use more adjectives of any kind when an object is part of a size contrast set, we also annotated responses for the presence of color adjectives. Note that, when a size contrast was present, the two objects in an image did not differ in color, so using color to refer to the target object would be over-informative.

We modeled adjective presence (a binary outcome variable) with a logistic mixed-effects model, using the lme4 package in R (Bates et al., 2015). The model included fixed-effects for contrast condition (size contrast present vs. absent) and adjective type (size vs. color), and interaction between them, and a random intercept by item. We predicted an interaction between the fixed effects, such that the presence of a size contrast would increase the use of size adjectives but not color adjectives. Following the interaction test, we conducted Bonferroni-corrected pairwise comparisons to test simple effects of contrast (present vs. absent) for size adjectives, and for color adjectives, using the multcomp package (Hothorn, Bretz, & Westfall, 2008).

Results

For GPT-4o, as predicted, we found a significant interaction between contrast and adjective type (β=0.88, SE=0.33, z=2.6, p=0.025). Simple effects tests revealed that the presence of a size contrast increased the use of size adjectives (β=0.95, SE=0.25, z=3.79, p<0.001) but not color adjectives (β=0.069, SE=0.22, z=0.31, p=1). Specifically, GPT-4o generated size adjectives on 64.41% of trials when a size contrast was present, but only on 25.86% of trials without a size contrast. It generated color adjectives on 52.54% of trials with a size contrast, and 50.00% of trials without a size contrast.

For LLaVA, the interaction was not significant (β= -0.028, SE=0.34, z=-0.083, p=1), and neither were the simple effects of the presence of a size contrast on the generation of size (β=0.20, SE=0.26, z=0.77, p=1.00) or color adjectives (β=0.23, SE=0.21, z=1.06, p=0.87). LLaVA generated size adjectives on 23.08% of trials featuring a size contrast versus 17.31% of trials with no contrast, and it generated color adjectives on 51.92% of trials featuring a size contrast versus 42.31% of no-contrast trials. We did not analyze data for InstructBLIP, because it did not generate size adjectives in either condition.

Discussion

Our findings provide evidence that vision-language transformer models exhibit multimodal integration of visual and linguistic context in a manner that resembles pragmatic inference. In Experiment 1, we analyzed how these models processed referring expressions as input. We found that LLaVA and InstructBLIP exhibited greater surprisal when such expressions were infelicitous (violating the maxim of quantity) vs. felicitous, given the context of the input image. Two analyses (1 and 4) demonstrated higher surprisal when a size adjective was *over*-informative (e.g., "*the big hammer*" for an image with a single hammer) than when it provided necessary disambiguating information (e.g., for an image with two hammers of different sizes). Analysis 3 provided evidence for higher surprisal when a size adjective was *under*-informative (e.g., "*the big cork*" in the presence of two equally sized corks) than when it supported disambiguation. Analysis 5 found evidence for pragmatic effects in the absence of size adjectives in LLaVA, but not in InstructBLIP, indicating that the effects we found in the former model are context-based and do not depend on lexical semantic information contained in size adjectives.

However, in yet another analysis (2), the results showed a pattern that was opposite of what would be expected from a pragmatically sensitive system. Surprisal was lower in an infelicitous, under-informative condition (e.g., "*the hammer*" for an image with two hammers), than in three felicitous conditions (e.g., for an image with a single hammer). This pattern may reflect an expectation for the text to refer to the most frequent object in the image: the object referred to by

the text was the most frequent one in the infelicitous condition, but not in the felicitous conditions. Therefore, object frequency may influence next-token predictions in vision-language transformers.

However, we observed these object-frequency effects only when one object was strictly more frequent than the others (Analysis 2), but not when different objects were equally frequent (Analysis 3). Thus, our results indicate that when object frequency and pragmatic felicity are explicitly pitted against one another, object frequency may dominate. (Another difference between those two analyses was that only Analysis 3, which did support pragmatic-like processing, included size adjectives; but, as described above, the presence of an adjective is likely not fully responsible for triggering such processing, as indicated by Analysis 5).

These findings suggest that pragmatic-like processes in multimodal transformers are less robust to the statistical features of their immediate context compared to pragmatic processes in humans. This hypothesis is consistent with other aspects of sensitivity to input statistics in LLMs, such as syntactic processes in text-only LLMs that appear more sensitive than humans to word frequency (e.g., Lasri et al., 2022; Wei et al., 2021). A notable difference, however, is that whereas previous studies have identified effects of frequency encountered during model pre-training (which influences model priors/weights), here putative frequency effects are hypothesized to be due to in-context frequency.

In Experiment 2, we analyzed how vision-language transformers generated referring expressions as output. We found that the responses of GPT-4o were sensitive to pragmatic factors: it produced size adjectives (e.g., "*the big hammer*") more often when the target object was part of a size contrast set than when it was not. This pattern reflected a pragmatic-like process, not a general tendency for elaborate descriptions of objects in size contrasts, because color adjectives were not produced more often when the target object was part of a size contrast. LLaVA also generated descriptively more size adjectives when the target object was part of a size contrast set, but this effect was not significant. InstructBLIP did not show evidence for pragmatic effects in the object description task, generating no size adjectives both when a size contrast was present and absent.

When testing LLaVA and InstructBLIP (but not GPT-4o), the prompt named the specific target object that the model should refer to (in addition to its spatial location). The reason is that, without such prompting, these two models frequently referred to an incorrect object, rendering trials invalid for analysis. One potential issue with our prompting technique is that the models could rely on a text-based copying heuristic to respond to the prompt without relying on visual context, even when the result is an under-informative expression (for an overview of such a mechanism—"copy attention heads"—in text-only LLMs, see Olsson et al., 2022; Crosbie & Shutova, 2025; Wang et al., 2022). A potential solution to this issue is to test models like LLaVA and InstructBLIP on simpler image displays (e.g., containing only two objects), which may increase the likelihood of correctly naming the object in the specified spatial location, even when the prompt does not explicitly name it. Additionally, the artificial stimuli used in this study may have degraded visual performance for LLaVA and InstructBLIP, so we plan to test these models using photo-realistic images.

Whereas our surprisal analyses in Experiment 1 provided evidence for pragmatic-like inference in LLaVA and InstructBLIP, the image description task in Experiment 2 did not. What may cause this disparity in results between the two experiments? On the one hand, model surprisal and generation abilities are two sides of the same coin: surprisal is derived from next-token probability, and next-token probability determines the output at each step during autoregressive inference (Brown et al., 2020). On the other hand, we did not analyze surprisal for the most probable next-token (which would necessarily be generated by the model when deterministic sampling is used); instead, we analyzed surprisal for the specific target words in our stimuli, which would not always be the most probable next-tokens. Therefore, pragmatic effects may be sufficient to influence the internal next-token probability *distributions* of LLaVA and InstructBLIP, but the effects may not be strong enough to influence the single *most probable* token at each step of text generation.

In summary, our study presents evidence for processes resembling multimodal pragmatic inference in three vision-language transformer models: LLaVA, InstructBLIP and GPT-4o. Whereas we believe our results provide strong support for the existence of pragmatic-like processes, they appear to be subtle in nature. First, they are not robust to the statistical properties of visual inputs, highlighting an important difference from pragmatic processing in humans. Second, pragmatic effects reflected by next-token probability distributions may be too weak to determine response generation, as least with deterministic response sampling. These qualifications notwithstanding, the ability of these models to combine information across input domains (visual and linguistic) to support pragmatic-like processes is an exciting example of situated language use that is key to human communication.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901.

Brown-Schmidt, S., & Konopka, A. E. (2011). Experimental Approaches to Referential Domains and the On-Line Processing of Referring Expressions in Unscripted Conversation. *Information*, *2*(2). https://doi.org/10.3390/info2020302

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*(4), 592–609. https://doi.org/10.1016/j.jml.2005.12.008

Chang, T. A., & Bergen, B. K. (2024). Language Model Behavior: A Comprehensive Survey. *Computational Linguistics*, *50*(1), 293–350. https://doi.org/10.1162/coli_a_00492

Cho, Y., & Kim, S. (2024). Pragmatic inference of scalar implicature by LLMs. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 10–20). https://doi.org/10.18653/v1/2024.acl-srw.2

Chronis, G., & Erk, K. (2020). When is a bishop not like a rook? When it`s like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 227–244). https://doi.org/10.18653/v1/2020.conll-1.17

Cong, Y. (2024). Manner implicatures in large language models. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-80571-3

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3), e13256. https://doi.org/10.1111/cogs.13256

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X

Crosbie, J., & Shutova, E. (2024). Induction heads as an essential mechanism for pattern matching in in-context learning. *arXiv preprint arXiv:2407.07011*.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 49250–49267.

Degen, J., & Tanenhaus, M. K. (2015). Processing Scalar Implicature: A Constraint-Based Approach. *Cognitive Science*, *39*(4), 667–710. https://doi.org/10.1111/cogs.12171

Do, M. L., Papafragou, A., & Trueswell, J. (2020). Cognitive and pragmatic factors in language production: Evidence from source-goal motion events. *Cognition*, *205*, 104447. https://doi.org/10.1016/j.cognition.2020.104447

Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. G. (2017). Reference Production as Search: The Impact of Domain Size on the Production of Distinguishing Descriptions. *Cognitive Science*, *41*, 1457–1492. https://doi.org/10.1111/cogs.12375

Gauthier, J., Hu, J., Wilcox, E., Qian, P., & Levy, R. (2020). SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 70–76). https://doi.org/10.18653/v1/2020.acl-demos.10

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences*, *7*(2), 92–96. https://doi.org/10.1016/S1364-6613(02)00040-2

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Grodner, D., & Sedivy, J. (2011). The effect of speaker-specific information on pragmatic inferences. *The processing and acquisition of reference*, *2327*, 239-72.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8. https://doi.org/10.3115/1073336.1073357

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641–671. https://doi.org/10.1037/bul0000145

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023). A fine-grained comparison of pragmatic language understanding in humans and language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4194–4213). https://doi.org/10.18653/v1/2023.acl-long.230

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A Systematic Assessment of Syntactic Generalization in Neural Language Models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744). https://doi.org/10.18653/v1/2020.acl-main.158

Ichien, N., Stamenković, Dušan, & and Holyoak, K. J. (2024). Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors. *Metaphor and Symbol*, *39*(4), 296–309. https://doi.org/10.1080/10926488.2024.2380348

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The Effect of Scene Variation on the Redundant Use of Color in Definite Reference. *Cognitive Science*, *37*(2), 395–411. https://doi.org/10.1111/cogs.12019

Lasri, K., Seminck, O., Lenci, A., & Poibeau, T. (2022). Subject Verb Agreement Error Patterns in Meaningless Sentences: Humans vs. BERT. *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 37–43). https://aclanthology.org/2022.coling-1.4/

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning*, *202*, 19730–19742.

Lin, F., Altshuler, D., & Pierrehumbert, J. B. (2024). Probing Large Language Models for Scalar Adjective Lexical Semantics and Scalar Diversity Pragmatics. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 13033–13049). https://aclanthology.org/2024.lrec-main.1141/

Linzen, T. (2020). How Can We Accelerate Progress Towards Human-like Linguistic Generalization? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 5210-5217).

Linzen, T., & Baroni, M. (2021). Syntactic Structure from Deep Learning. *Annual Review of Linguistics*, *7*, 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, *36*, 34892–34916.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

Noveck, I. A., & Reboul, A. (2008). Experimental Pragmatics: A Gricean turn in the study of language. *Trends in Cognitive Sciences*, *12*(11), 425–431. https://doi.org/10.1016/j.tics.2008.07.009

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., … Olah, C. (2022). In-context Learning and Induction Heads. *arXiv preprint arXiv:2209.11895*. https://doi.org/10.48550/arXiv.2209.11895

OpenAI (2023). OpenAI – API reference

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., … Malkov, Y. (2024). GPT-4o System Card. *arXiv preprint arXiv:2410.21276*. https://doi.org/10.48550/arXiv.2410.21276

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, *86*(3), 253–282. https://doi.org/10.1016/S0010-0277(02)00179-8

Petersen, E., & Potts, C. (2023). Lexical Semantics with Large Language Models: A Case Study of English "break." *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 490–511). https://doi.org/10.18653/v1/2023.findings-eacl.36

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. In *From fieldwork to linguistic theory: A tribute to Dan Everett* (pp. 353–414).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rubio-Fernandez, P., Berke, M. D., & Jara-Ettinger, J. (2025). Tracking minds in communication. *Trends in Cognitive Sciences*, *29*(3), 269–281. https://doi.org/10.1016/j.tics.2024.11.005

Ruis, L. E., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2022). Large language models are not zero-shot communicators.

Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information Integration in Modulation of Pragmatic Inferences During Online Language Comprehension. *Cognitive Science*, *43*(8), e12769. https://doi.org/10.1111/cogs.12769

Sedivy, J. C. (2003). Pragmatic Versus Form-Based Accounts of Referential Contrast: Evidence for Effects of Informativity Expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23. https://doi.org/10.1023/A:1021928914454

Sedivy, J. C. (2004). Evaluating Explanations for Referential Context Effects: Evidence for Gricean Mechanisms in Online Language Interpretation. In *J. Trueswell & M. Tanenhaus (Eds.), Approaches to studying world-situated language use* (pp. 345–364). MIT Press.

Sedivy, J. C., K. Tanenhaus, M., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. https://doi.org/10.1016/S0010-0277(99)00025-6

Stowe, K., Utama, P., & Gurevych, I. (2022). IMPLI: Investigating NLI Models' Performance on Figurative Language. *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers) (pp. 5375–5388). https://doi.org/10.18653/v1/2022.acl-long.369

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, *268*(5217), 1632–1634. https://doi.org/10.1126/science.7777863

Tong, X., Shutova, E., & Lewis, M. (2021). Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4673–4686). https://doi.org/10.18653/v1/2021.naacl-main.372

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460. https://doi.org/10.1093/mind/LIX.236.433

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*. https://doi.org/10.48550/arXiv.1804.07461

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Wei, J., Garrette, D., Linzen, T., & Pavlick, E. (2021). Frequency Effects on Syntactic Rule Learning in Transformers. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 932–948). https://doi.org/10.18653/v1/2021.emnlp-main.72

Wilcox, E., Vani, P., & Levy, R. (2021). A Targeted Assessment of Incremental Processing in Neural Language Models and Humans. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 939–952). https://doi.org/10.18653/v1/2021.acl-long.76