Measuring the Effects of Visual Salience in Human and AI Descriptions with Image Editing

Nina Gregorio[1], **Edoardo M. Ponti**[1], **Sharon Goldwater**[1]
[1]Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
n.gregorio@sms.ed.ac.uk

Abstract

How does our perception of the world influence the way we talk about it? Psycholinguistic studies have investigated whether visual salience correlates with entity mention and ordering, but often disregarded its effect on grammar or relied on simplistic images or artificial cues. In this study, we explore the use of generative AI to better control for salience in visual stimuli while keeping them realistic, and to serve as a proxy for human participants in studying how different types of salience impact image descriptions. We consider three salience types: *perceptual* (e.g. relative size in the image), *inherent* (e.g. animacy), and *relational* (e.g. human–object interaction). We first analyze human- and AI-generated captions for natural images to examine how salience correlates with how early, and in what grammatical role, an entity is mentioned. We find strong correlations between models and humans in this observational study, justifying the use of AI models alone in a further causal study. For this second study, we created datasets composed of pairs of images, where we used an image-editing model to intervene on the salience of a target entity. We show that relational and perceptual salience lead to the entity being mentioned earlier in captions and being mapped to more prominent grammatical roles. The magnitude of this effect varies across entity types, with animate entities (high inherent salience) showing a particularly distinct pattern.

1 Introduction

Language is often used to communicate about the world, and an important question in psycholinguistics is how our perception of the world influences the way we talk about it. Researchers have made some progress in answering this question, for example by showing that visually salient entities are more likely to be mentioned and are mentioned earlier, when speakers describe a scene (Clarke et al., 2013, 2015; Spain and Perona, 2008, 2011; Barker et al., 2023). These studies rely on a wide range of measures to quantify salience, from eye-gaze measures (Griffin and Bock, 2000; Bock et al., 2013; Coco and Keller, 2012), low-level perceptual features such as brightness, position or relative size in the image (Spain and Perona, 2008, 2011; Clarke et al., 2015; Barker et al., 2023), to more conceptual features such as affordances or meaning (Henderson et al., 2018; Barker et al., 2023). A few studies have even looked at the relative impact of different types of salience on language production (Clarke et al., 2013; Henderson et al., 2018; Barker et al., 2023). However, most studies relating visual salience to language production have only examined surface-level features of the language (i.e., the order in which entities are mentioned; Griffin and Bock, 2000; Barker et al., 2023), and those that have analyzed syntactic choices have used simplified images (Griffin and Bock, 2000; Gleitman et al., 2007; Myachykov et al., 2012) and/or artificial salience interventions such as adding a subliminal flash (Gleitman et al., 2007), word cues (Coco and Keller, 2012), visual cues (Tomlin, 1995) or visual priming (Myachykov et al., 2012). These choices are partly due to the difficulty of creating the tightly controlled stimuli that are often used in language production experiments.

In this study, we consider how recent advances in generative AI may help address these issues by enabling the creation of more realistic stimuli for controlled experiments, and/or as proxies for human participants in an image description task. We consider three different overarching types of salience, inspired by previous work:

Perceptual salience: Based on low-level features that attract early eye gaze, such as relative size in the picture, depth, texture, brightness, or distance to the image center.
Inherent salience: Based on intrinsic prop-

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 301–317

(a) **H–O Interaction**

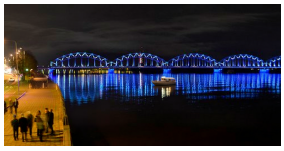
A *woman* is arranging a **fruit** *bowl* on the *kitchen table*, which includes *bananas* and *oranges*.

(b) **No H–O Interaction**

A *man* stands in a *kitchen*, looking towards the *stove* area, with a *bowl* of **fruit** sitting on the dining *table* in front of him.

(c) ↑ **Perceptual Salience**

A **boat** glides across the *water*, reflecting vibrant blue *lights* of a *bridge* that spans the *river*.



(d) ↓ **Perceptual Salience**

A group of *people* stand on a *waterfront promenade*, gazing at the illuminated blue *bridge*.

Figure 1: We created two datasets of image pairs where: 1) **Left**: We intervene on the relational salience of a target entity by adding a human interacting or not interacting with it; 2) **Right**: We intervene on the perceptual salience of a target entity by changing its size, centrality and/or depth. We then generated captions using different vision–language models (VLMs; e.g., LlaVa in these examples). Nouns are shown in italic, and the target entity (if mentioned) is in bold.

erties of entities (i.e., features that require entity identification), such as animacy or affordances.

Relational salience: Context-dependent salience based on how an entity interacts with other entities in the image.

We analyze how these different types of salience impact the captions generated by humans and vision-and-language models (VLMs).

We begin with an observational study comparing VLM and human captions on the COCO dataset (Lin et al., 2014). Here we use *size* as a measure of perceptual salience and *semantic category* (operationalized using the COCO labels, e.g., humans, animals, furniture, etc.) as a measure of inherent salience. We show that, despite differences in the overall caption length and number of entities mentioned, humans and VLMs follow similar patterns in terms of how these measures of salience correlate with linguistic outcomes (whether an entity will be mentioned, and if so, in what order and syntactic role).

The similarity between VLM and human results in this observational study suggests that VLMs may be useful for hypothesis generation or as pilot "participants" for further human studies. We illustrate this idea in our second study, where we demonstrate that both perceptual and relational salience causally affect image descriptions generated by VLMs, and present detailed results that suggest avenues for future human studies. This study also highlights the use of VLMs to create controlled but realistic stimuli through image-editing.

Specifically, we use an image-editing VLM to create two new datasets based on COCO, each consisting of image pairs that vary either the perceptual or relational salience of a target entity, without changing other aspects of the scene. We use our two datasets to explore the impact of different salience types on the content, linear order, and syntactic realization of scene descriptions generated by VLMs. In these studies, we find that both perceptual and relational salience raise the probability of being mentioned, the order of mention, and the grammatical role (according to the following grammatical hierarchy: subject > direct object > object of a preposition). We further show that the impact of perceptual and relational salience highly varies depending on the inherent salience of the target entity: Perceptual salience mostly influences linguistic properties of animate entities and relational salience mostly raises the order of mention of entities with low inherent salience.

Overall, our work sheds light on how different types of salience interact in image descriptions given by both humans and VLMs, and also demonstrates the promise of recent AI tools for studying these questions more deeply in future.[1]

2 Background

Multiple studies have investigated the impact of visual salience on language production, using tasks such as object naming (Clarke et al., 2013), referring expression generation (Clarke et al., 2015), or

[1]Our datasets are available at https://huggingface.co/collections/EdinburghNLP/salience-interventions, models and code at https://github.com/Naiina/salience-interventions.

full scene description (Spain and Perona, 2008; Gleitman et al., 2007; Coco and Keller, 2012; Barker et al., 2023; Henderson et al., 2018). However, defining and extracting salience features is itself a challenge. Several studies focused on low-level perceptual features such as relative size or distance to the center (Spain and Perona, 2008, 2011; Clarke et al., 2015; Barker et al., 2023), as such features have been shown to attract eye gaze (Underwood and Foulsham, 2006), and eye gaze in turn has been shown to predict entity mention (Griffin and Bock, 2000; Coco and Keller, 2012). More recent work has emphasized the significant role of high-level features, such as semantic content, informativity, and graspability, and has shown that these are more explanatory than low-level features in predicting eye gaze and the order in which entities are mentioned (Henderson et al., 2019; Barker et al., 2023). However, these high-level features are mapped out using subjective human ratings of picture patches. As well as being costly, such "meaningfulness" ratings are subjective and could be culturally specific. On the one hand, collecting such ratings across different cultures can evaluate whether meaningfulness is a universally predictive feature. On the other hand, using more objective and automatically obtainable high-level features (such as entity categories) would make it easier to scale studies to many languages/cultures, and would also allow researchers to examine cross-cultural differences in how these features impact language production.

While there has been considerable investigation of different salience features, most prior work has focused on fairly simple measures of their impact on language, such as whether an entity is mentioned, and when (i.e., in what order relative to other mentioned entities; Clarke et al., 2013, 2015; Spain and Perona, 2008, 2011; Barker et al., 2023). However, a few studies have also shown effects on more complex linguistic outcomes, for example by showing that artificially directing the attention of the speaker toward the agent or the patient of a simple scene influences the choice of a passive or active main verb (Gleitman et al., 2007; Myachykov et al., 2012). Finally, Berger et al. (2023) used a classifier to predict a wider range of linguistic features of the caption (negation, voice, numerals, transitivity) using the entire image. This demonstrates that features of a picture impact various linguistic properties, but it remains unclear *which* features they might be.

This previous work highlights several current challenges and gaps in the field: (1) There are several types of salience that interact with sentence production, but these are correlated (as we show in Section 3), which can make it difficult to tease apart their effects. (2) Studies manipulating salience directly did so using artificial cues and/or simplified drawings, probably due to the difficulty of manipulating realistic images, and the challenge of observing an effect of salience on grammar on small datasets and complex scenes. (3) Extracting high-level salience features currently relies on human subjective annotations. (4) To our best knowledge, work comparing the impact of several types of salience on language has focused on mention order only. (5) Most studies focused on English.

In this paper, we argue that VLMs can potentially address many of these limitations. First, image segmentation and entity classification can automatically extract features on entity size and category (as a measure of inherent salience). We demonstrate that these features can provide meaningful insights in both observational and causal studies.[2] Secondly, in our causal study, we used an image-editing model to intervene on salience features of entities in the COCO dataset (Lin et al., 2014), enabling us to isolate the impact of different types of salience features in realistic pictures, with minimal human annotation. Third, we then automatically generated captions of those pictures, enabling us to investigate salience effects on linguistic choices at a large data scale. Finally, although the experiments presented here are on English, the pipeline we developed would easily scale to multiple languages, and more culturally-diverse pictures, which would allow cross-linguistic and cross-cultural comparisons—areas we intend to explore in future work.

3 Comparing AI and Human Captions on Natural Images

In this section, we use mixed-effects models to analyze both human and AI captions on unmodified pictures from the COCO dataset. We use relative size as a measure of perceptual salience, and COCO super-category label as a measure of inherent salience. Using the human captions, we (1) confirm that these measures correlate with both probability of mention and order of mention, as pre-

[2] Although the COCO segmentations are hand-annotated, VLMs could provide similar information for other datasets.

dicted by previous studies; and (2) demonstrate that these measures also correlate with the probability of being the subject of a sentence (not covered by previous work to our best knowledge). In addition, we show that (3) these correlations are also present in the AI captions; and (4) although the caption length and number of mentions are much higher for AI captions than for human ones, the effects of our perceptual and inherent salience measures are similar in both cases (despite small but statistically significant differences).

3.1 Method

Image dataset and salience features We use the COCO dataset, which provides segmentation for entities of 12 super-categories (for a full list, see Appendix A). We consider 4913 images sampled from the validation dataset from 2014 and 2017. We use the COCO-segmented semantic categories as proxies for inherent salience. Entities labeled as persons and animals are considered to have high inherent salience due to animacy; we assume that other categories will have various inherent salience, although our study is not designed to identify those latent values. For each segmented entity, we compute its relative size (pixel area as a proportion of image size) as a measure of perceptual salience.

Human and AI-generated Captions We compare the human captions provided as part of COCO to AI-generated captions from three state-of-the-art open-source VLMs: LLaVA-OneVision-1.5 (An et al., 2025), InternVL3 (Zhu et al., 2025), and Qwen2.5-VL (Bai et al., 2025). For details of the prompts and generation procedure, see Appendix F.

Linguistic Feature Extraction We identify the nouns in each caption with spaCy,[3] extract their grammatical role, and compute their relative word order. Further details are available in Appendix B.

Entity-to-noun mapping For each segmented entity, we need to identify the corresponding noun in the caption, if any. As a first step, we look for a perfect match between the entity's COCO category name and the nouns in the sentence. If this is not found, we use Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) to determine whether the target entity is present in the caption (see prompt in Appendix E). If so, the model was prompted to return the first mention of the corresponding noun.

[3] https://spacy.io/

Caption source	Avg # words	Avg # nouns	Avg # subjs	Avg # mentions	Mention rate
Human	10.41	3.69	0.40	1.31	0.55
Qwen	27.00	8.62	1.41	1.69	0.62
InternVL	25.24	7.89	1.36	1.54	0.65
LLaVA	24.90	7.81	1.49	1.70	0.64

Table 1: Statistics of human and AI captions, including the average number of words, nouns, and grammatical subjects in captions from different sources, as well as the number of mentions (nouns that correspond to a COCO segmented entity category), and the average proportion of categories present in each image that are mentioned.

To check the reliability of our pipeline, we randomly sampled 20 examples per category (240 in total) and manually annotated them. Our pipeline reached 98% accuracy on this subset.[4]

Dataset statistics The average number of distinct COCO categories of segmented entities per image is 2.9. Table 1 shows the average number of words, nouns, subjects, and mentions in the human and AI captions. These statistics indicate that the VLMs generate longer captions than humans overall, with other statistics correspondingly higher as well. Despite these differences, the following section shows that the relationship between our salience measures and linguistic outcomes is strikingly similar between the VLMs and humans.

Statistical analysis We fit mixed-effect regression models to predict 1) whether an entity is mentioned (logistic regression), and if so, 2) the normalized mention order (rank of the target noun with respect to other nouns, normalized by the length of the sentence to account for the overall length differences between human and AI; linear regression), and 3) whether it is a grammatical subject (logistic regression). All three models include log-size, semantic super-category, and caption source (human and each of three VLMs) as fixed effects, and a random intercept per image. We also include interaction terms between the source and the other fixed effects, to assess whether the different types of salience have different effects on captions generated by humans vs. models. More details about the models are available in Appendix C.

[4] Since our pipeline does not determine which instance of an entity the noun refers to, when many of the same category are present, we compute size measures as the sum of the sizes of all entities of the given category. This limitation is addressed in the second study, where we select only images with the target entity appearing once in the picture to ensure a unique noun–entity mapping.

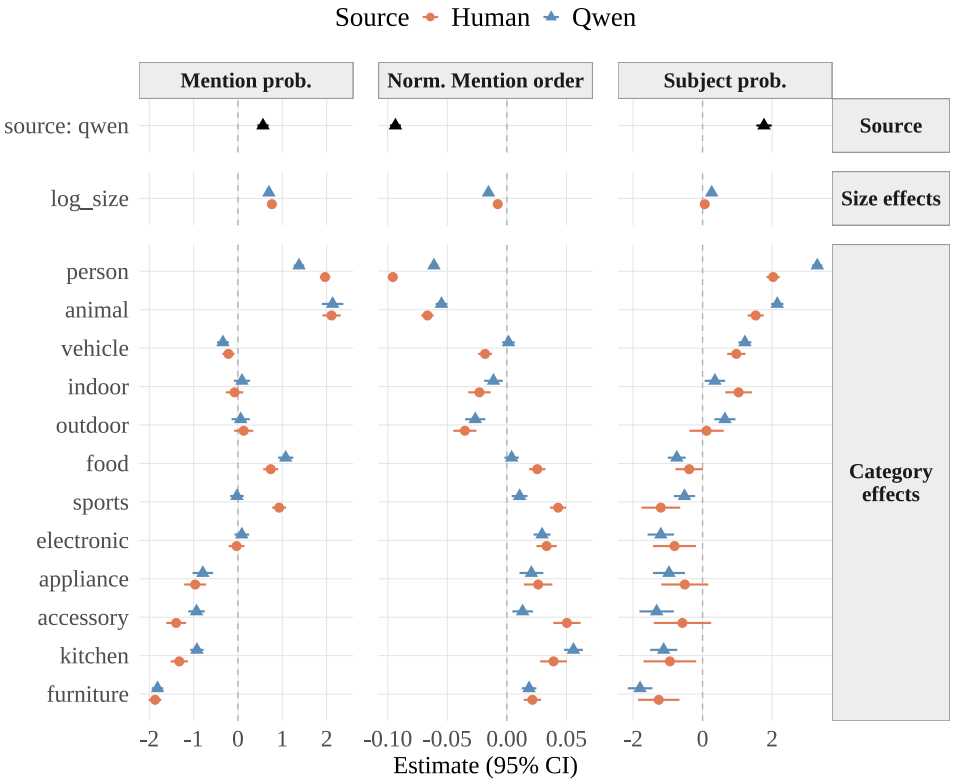


Figure 2: Effects of source, size and category in mixed-effects models fitted to predict mention probability, normalized mention order, and subject probability of entities. The category effects are plotted against the source-specific grand means to facilitate comparison between sources, and show that although the differences between AI and human for each category are often significant (all p-values available in Appendix C, Table 3), these source-interaction effects are typically much smaller than the main effect of that category.

To check that the model coefficients are identifiable, we assessed collinearity between size and category by fitting a linear regression model to predict size from category. We found an R^2 of 0.1715, indicating a low (though highly statistically significant: $p < 2.2\text{e-}16$) degree of collinearity, i.e., fairly reliable coefficients. We also report model comparisons to further support our main claims in Table 2. Specifically, for each outcome measure, we performed ablations with respect to the full mixed effects models, leaving out either the source interaction terms (*No source interaction*), the main effects and interactions for size (*No size*), or category (*No category*). The Akaike Information Criterion differences (ΔAIC) are computed between each ablation and the full model (all trained on the entire dataset). The accuracy and MSE scores are computed on models trained on 70% of the dataset (random split) and evaluated on the other 30%. Random effects are set to zero during prediction.

	Model	Accuracy	MSE	ΔAIC
Mention	Full model	0.756	—	0
	No src interact	0.756	—	162
	No size	0.685	—	8949
	No category	0.700	—	7901
Rank	Full model	—	0.0062	0
	No src interact	—	0.0063	476
	No Size	—	0.0066	1919
	No category	—	0.0080	6865
Subject	Full model	0.780	—	0
	No src interact	0.781	—	244
	No size	0.772	—	234
	No category	0.621	—	8478

Table 2: Comparisons between the full models and ablations that remove either source interactions, size, or category. We report accuracy (for mention and subject) and MSE (for rank), as well as ΔAIC.

3.2 Results and Discussion

Overview In Figure 2, we present the coefficients of the three mixed effects models. For clarity, only the Qwen VLM is shown, because coefficients for the other VLMs are similar (see Appendix C, Figure 5). Figure 2 shows that both size and category have significant (often large) effects on predicting linguistic outcomes for both AI and humans. Furthermore, despite an overall average shift in AI

linguistic outcomes compared to human captions (explained by AI's higher verbosity and indicated by the non-zero source coefficients), the source-interaction effects are small compared to the other fixed effects—that is, size and category affect linguistic outcomes in AI and humans in mostly similar ways. This conclusion is further supported by the comparisons in Table 2, which show that removing the source-interaction effects has less impact than removing the other fixed effects (especially Category).

Source fixed effects and AI verbosity As expected from Table 1, our statistical models find that AI is more likely than humans to mention entities and to use them as subjects (AI Source terms > 0 for Mention prob. and Subject prob.). The AI Source term for mention order is < 0 because AI captions are longer on average, which lowers the normalized rank of words appearing early in the sentence.

Size Size has a clear effect on mention probability, and a smaller but still significant effect on mention order and subject probability (both conditioned on the entity being mentioned at all). Ablation studies nevertheless show a drop in performance and a higher AIC for all models without size terms, indicating that size, and thus perceptual salience, contributes to predicting grammar features of the nouns in the caption. We might expect larger effects of perceptual salience if we included additional (or more sophisticated) perceptual features.

Category Ablation studies show that the drop in performance and AIC increase is the highest for models without category fixed effects, confirming that category, and thus inherent salience, strongly predicts linguistic outcomes. Moreover, we observe overall similar patterns between AI and humans. Especially, for a given size, animals and persons tend to be mentioned more and earlier than other categories, and are more likely to be subjects. Additionally, two inanimate categories are likely to be subjects: vehicles, potentially due to their mobility, and indoor entities, which would require further investigation.

Source interaction terms Although models without these interactions have a significantly worse fit according to AIC, they show little to no performance drop compared to the full models. The size–source interaction term are small compared to the other model coefficients, indicating that the effect of size differs only slightly between humans and AI. Similarly, the difference between human and AI category coefficients is significant for half of the coefficients for mention probability and subject probability, and for most of them for mention order (see Table 3 in Appendix C). But the interaction terms remain small compared to the absolute category effects.

Together, those observations indicate that most of the predictive power of the mixed-effect model is carried by the main effects of size and category, and varies little with respect to source. This justifies the use of AI-only captions in the second experiment.

4 Causal studies: How salience types affect AI-generated captions

4.1 Overview

The previous study provided evidence that perceptual and inherent salience are independently correlated with the linguistic form of image descriptions. In this section, we perform a causal study on perceptual and relational salience to demonstrate that for VLMs, each of these is not just correlated with linguistic features, but causally affects them. We create datasets that intervene on one of these two manipulable types of salience, and study how that affects grammar features (mention, word order and grammatical role) of descriptions of images.[5]

Image and Target Entity Selection To enable us to change one entity without completely changing the focus of the picture, we filtered out COCO images containing fewer than 5 segmented entities or occupied for more than 50% by a single segmented entity type. For each image, we selected at random a target entity that satisfies the two following conditions. The target entity's category must be unique, to ensure an unambiguous mapping between it and the noun referring to it. Moreover, we only consider entities of medium size, i.e., covering between 1% and 25% of the image, so that the entity can be modified without completely changing the nature of the scene. Note that these selection criteria shift the overall distribution of the target entities compared to the initial entity distribution in COCO (see Figure 7 in Appendix G).

Interventions For each target entity, we intervene in order to vary its perceptual or relational salience. This results in pairs of images, where the

[5]Intervening on inherent salience is more challenging, as it might lead to incoherence in the scene (e.g. a car indoors).

target entity has low and high salience, respectively, while the rest of the picture is kept constant. We operationalize the intervention through the following features: relative size, distance, and depth for perceptual salience; and Human-Object Interaction (HOI) for relational salience. These features were chosen based on prior evidence of their correlation with language production (see Section 2).

Next, each picture was captioned by the same three VLMs used in our first study, yielding a **Perceptual Dataset** (1042 pairs of pictures) and an **HOI Dataset** (910 pairs of pictures), as illustrated in Figure 1. We used these datasets to investigate how the variation of each type of salience impacts linguistic outcomes and how these salience types interact.

4.2 Details of Dataset Creation

Perceptual Salience Pairs We intervened on the salience features of an entity already present in the picture, rather than adding a new entity, in order to ensure that the target entity is coherent with the rest of the scene. We used Qwen-Image-Edit-2509 (Wu et al., 2025), a state-of-the-art open source editing model, to modify images from the COCO dataset. We found the model performed better at adding entities than modifying existing ones, so we first asked the model to remove the target entity, then created two new images by adding it back. First we simply prompted to add the entity, which by default appears big, central, and in the foreground. We then prompted the model to add a target entity that is "*small or in the background*" (see prompts in Appendix D). We did not add any constraint on position, as there were often only a few positions available that would keep the picture realistic in terms of perspective and relation to other entities.

We manually filtered pictures, deleting pictures containing entities that were incomplete, floating, duplicated, in a surprising place, or with unrealistic size. We further excluded pictures with modifications other than the target entity, and those in which the perceptual salience of the target entities did not visibly differ in at least one of the low-level salience features (relative size, depth, centrality). We also excluded pictures with a high color contrast difference between the two target entities because it has been argued that color also impacts salience. Examples of pictures we manually filtered out are available in Appendix G.

Human–Object Interaction Pairs We considered only pictures with no human originally present in the picture, to ensure that no human is already interacting with the target entity. We prompted Qwen-Image-Edit-2509 to generate pairs of pictures, adding 1) a human interacting with the target entity and 2) a human standing on the side (see prompts in Appendix D). We considered as valid "interaction" pictures where the person is looking at, pointing at, or touching the target entity. We manually filtered out pictures with unclear interaction, with the human interacting with the wrong entity or hiding the target entity, as well as pictures with other modifications of the rest of the scene.

Target Entity Distribution Qwen-Image-Edit-2509 is more likely to generate a realistic picture for some entities than others. Unrealistic depictions include floating food, duplicated backpacks, etc. After automatic picture selection and manual filtering, the distribution of target entities differs from the original COCO entity distribution, as Figure 7 in Appendix G illustrates.

4.3 Results and Discussion

Figure 3 shows the main results for this section, with outcomes aggregated over the captions generated by all three VLMs, since different models behaved similarly. Plots for individual models can be found in Appendix H.

Probability of Mention Overall, the mention probability is greater for entities with high (82%) vs. low (63%) perceptual salience ($\chi^2(1) = 290$, $p < 0.001$), evidence that the correlation found in the first study is in fact causal. We also found an effect for relational salience (high: 78%; low: 62%; $\chi^2(1) = 170$, $p < 0.001$).

Breaking down these results per category, Figure 3 shows that both types of salience significantly influence the probability of mention across most categories. Animals are a clear exception: for these images, they are almost always mentioned regardless of their salience. Humans also have a very high probability of mention (consistent with the observational study), but mentions are still sensitive to perceptual salience (by definition there are no human target entities in the relational dataset).

Mention Order In both datasets, high-salience entities are introduced significantly earlier in captions (significant difference in both cases, using Wilcoxon signed-rank test with p < 0.001). In the

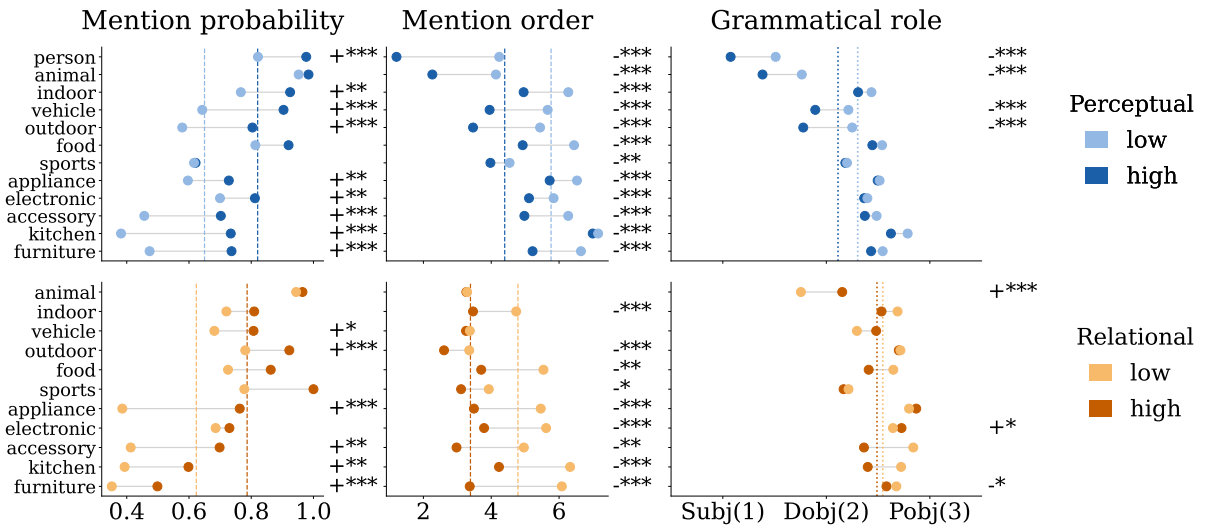


Figure 3: Effects of perceptual salience (top row, blue) and relational salience (bottom row, brown) on mention probability, mention order, and grammatical role, with results aggregated over all three VLMs. Dots show results for each category, and dotted lines show averages across all categories. To the right of each plot are significance values (chi-square test for Mention probability, and Wilcoxon signed-rank for Mention order and Grammatical role: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$). +/- shows the direction of significant effects. Grammatical roles are treated as numerical values (1, 2, 3) for visualization, but as an ordinal variable in significance tests.

perceptual dataset, the average rank is 4.1 for high-saliency vs. 5.1 for low-saliency entities. In the relational saliency dataset, the average rank is 3.2 for high-saliency vs. 4.5 for low-saliency entities.

However, perceptual and relational salience influence word order differently depending on the entity category. Specifically, human–object interaction has no significant effect on the average order of mention for animate entities (e.g., animals) or dynamic entities (e.g. vehicles). In contrast, in the perceptual salience dataset, animate and dynamic entities exhibit the largest average word order difference (e.g., -3.3 for humans, -2.2 for vehicles, -1.9 for animals). Although animate entities tend to appear early within sentences in the absence of visual stimuli (Gregorio et al., 2025), this work shows that the order of mention of animates is highly modulated by perceptual salience.

Grammatical Role In this subsection, we test whether greater salience leads to nouns being mapped to higher grammatical functions according to the following hierarchy: subject (S) > direct object (DO) > object of a preposition (PO). Concept-to-grammatical role mapping in language production has been widely theorized (Bock and Warren, 1985; Bock and Levelt, 1994), but to our knowledge, only a few studies have investigated grammar structures and visual perception (mostly looking at voice), and such studies used synthetic data or artificial cues (Tomlin, 1995; Gleitman et al., 2007).

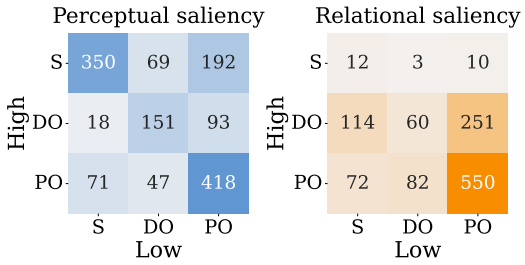


Figure 4: Counts of the grammatical roles occupied by target entities in image pairs from the perceptual (left) and relational (right) datasets. Rows and columns indicate the roles occupied when the entity has high or low salience, respectively.

In both datasets, when comparing the pairs of grammatical roles assigned to entities of high and low salience (respectively rows and columns in Figure 4), we observe a plurality of pairs where both target nouns are objects of a preposition: 28.8% in the perceptual dataset and 47.6% in the relational dataset. This is coherent with the frequent use of locatives in picture descriptions (e.g. "*next to the couch*", etc.). More generally, the majority of entity mentions share the same grammatical role in both datasets (diagonal cells): 63.3% for perceptual salience and 53.8% for relational salience.

In the perceptual salience dataset (Figure 4 left), the main grammatical role shifts (off-diagonal cells) that occur when decreasing salience are from subject to object of a preposition (13.2%) and di-

rect object to object of a preposition (6.4%). This is consistent with the hypothesis that reducing perceptual salience leads to the target nouns being mapped to lower grammatical functions. Figure 3 shows that perceptual salience significantly impacts the overall grammatical role distribution (Wilcoxon signed-rank test with $p < 0.001$). Breaking down per category, the grammatical role of humans, animals, and vehicles is significantly shifted, which is consistent with the large effects found for the subject probability of these categories in Section 3. Surprisingly, the outdoor category (which includes, e.g., traffic lights and benches) is also significantly impacted.

As for the relational dataset (Figure 4 right), results show that target entities are almost never the subject when a human is interacting with them: in almost all captions of pictures with HOI, the subject position is occupied by the human, preventing the target entity from accessing this grammatical role. Moreover, the main grammatical role shift is from direct object to object-of-a-preposition (21.7%), which is coherent with the hypothesis that reducing salience moves entities to lower grammatical roles. The second most common shift is from direct object to subject (9.8%). Figure 3 shows that this mostly concerns animals, which show a significant tendency to occupy higher roles (i.e., subject role) without HOI, while other categories show no significant effect or significant but very small effects. As a result, a Wilcoxon signed-rank test performed on all categories indicates that the overall shift is not significant.

5 Discussion and Conclusions

In this work, we investigated how several types of visual salience impact linguistic features in picture descriptions. We first showed that AI-generated and human-written captions exhibit similar associations between perceptual salience (measured as relative size), inherent salience (measured as entity category), and features of linguistic descriptions (content, order, and subjecthood). This similarity justifies the use of AI captions in our subsequent study. Although parts of COCO are likely included in the VLMs' training data, overfitting to this particular data set is not likely to explain the similar patterning with respect to salience, since the VLM captions deviate considerably from the human ones in both overall length and number of entities mentioned.

To isolate the causal impact of different types of salience on language, we then created two datasets, intervening on the relational and perceptual salience of target entities while leaving everything else unchanged. We showed that in AI-generated captions, increasing either perceptual or relational salience of a target entity makes it more likely to be mentioned, and mentioned earlier (consistent with previous studies on human data), and also makes it more likely to take a prominent grammatical role. However, the impact of salience changes is highly dependent on the inherent salience of the target entity. Animals are highly mentioned overall, regardless of their relative or inherent salience. In terms of mention order, relational salience has little to no effect for animate entities (animals) and dynamic entities (vehicles), whereas those are the categories where perceptual salience modulates mention order the most. For grammatical role, both types of salience have the largest impact on animate entities.

We do not suggest that VLMs can answer *how* visual salience impacts human utterance planning and production at the process level, nor is it clear whether the similarities we found between humans and VLMs would extend to other types of visually prompted tasks, such as generating referring expressions. Nevertheless, our results are exciting because they suggest that VLMs may provide a good model of human *behavior* in image description tasks, showing how different types of salience impact both linearization and syntactic choices. AI can also be useful for producing controlled but realistic image pairs for human experiments.

Together, these observations pave the way for broader investigation. For example, word order and grammatical role are highly correlated in English; future work should investigate languages with more flexible word order to disentangle the impact of salience on each of these linguistic features. Moreover, the current COCO dataset lacks scene type and cultural diversity, mostly representing Western-centric interior and exterior scenes. Culture influences what people mention when describing pictures (Berger and Ponti, 2025), and could similarly affect mention order and grammatical role. More generally, we hope that this work will inspire more systematic use of AI tools for both stimuli creation and to obtain fine-grained insight into complex interactions between language and visual perception.

Acknowledgements

We would like to sincerely thank Coleman Haley, Jennifer Culbertson, Frank Keller, Uri Berger and Hannah Rohde for their valuable feedback, as well as all the anonymous reviewers for their comments.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. 2025. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL technical report.

M Barker, G Rehrig, and F Ferreira. 2023. Speakers prioritise affordance-based object semantics in scene descriptions. *Language, cognition and neuroscience*, 38(8):1045–1067.

Uri Berger, Lea Frermann, Gabriel Stanovsky, and Omri Abend. 2023. A large-scale multilingual study of visual constraints on linguistic selection of descriptions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2285–2299, Dubrovnik, Croatia. Association for Computational Linguistics.

Uri Berger and Edoardo Ponti. 2025. Cross-lingual and cross-cultural variation in image descriptions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9453–9465.

J Kathryn Bock and Richard K Warren. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1):47–67.

Kathryn Bock, David E Irwin, and Douglas J Davidson. 2013. Putting first things first. In *The interface of language, vision, and action*, pages 249–278. Psychology Press.

Kathryn Bock and Willem JM Levelt. 1994. Language production: Grammatical encoding. In *Handbook of psycholinguistics*, pages 945–984. Academic Press.

Alasdair D. F. Clarke, Micha Elsner, and Hannah Rohde. 2015. Giving good directions: Order of mention reflects visual salience. *Frontiers in Psychology*, Volume 6 - 2015.

Alasdair DF Clarke, Moreno I Coco, and Frank Keller. 2013. The impact of attentional, linguistic, and visual features during object naming. *Frontiers in Psychology*, 4:927.

Moreno I Coco and Frank Keller. 2012. Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive science*, 36(7):1204–1223.

Lila R Gleitman, David January, Rebecca Nappa, and John C Trueswell. 2007. On the give and take between event apprehension and utterance formulation. *Journal of memory and language*, 57(4):544–569.

Nina Gregorio, Matteo Gay, Sharon Goldwater, and Edoardo Ponti. 2025. The cross-linguistic role of animacy in grammar structures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7349–7363.

Zenzi M Griffin and Kathryn Bock. 2000. What the eyes say about speaking. *Psychological science*, 11(4):274–279.

John M Henderson, Taylor R Hayes, Candace E Peacock, and Gwendolyn Rehrig. 2019. Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision*, 3(2):19.

John M Henderson, Taylor R Hayes, Gwendolyn Rehrig, and Fernanda Ferreira. 2018. Meaning guides attention during real-world scene description. *Scientific reports*, 8(1):13504.

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 *European conference on computer vision*, pages 740–755. Springer.

Andriy Myachykov, Simon Garrod, and Christoph Scheepers. 2012. Determinants of structural choice in visually situated sentence production. *Acta psychologica*, 141(3):304–315.

Merrielle Spain and Pietro Perona. 2008. Some objects are more equal than others: Measuring and predicting importance. In *European conference on computer vision*, pages 523–536. Springer.

Merrielle Spain and Pietro Perona. 2011. Measuring and predicting object importance. *International Journal of Computer Vision*, 91(1):59–76.

Russell S Tomlin. 1995. Focal attention, voice, and word order. *Word order in discourse*, pages 517–552.

Geoffrey Underwood and Tom Foulsham. 2006. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, 59(11):1931–1949.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025. Qwen-Image technical report.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models.

A COCO categories and super-categories

The list of categories and super-categories segmented in the COCO dataset is as follows:

- person: person
- vehicle: bicycle, car, motorcycle, airplane, bus, train, truck, boat
- outdoor: traffic light, fire hydrant, stop sign, parking meter, bench
- animal: bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe
- accessory: backpack, umbrella, handbag, tie, suitcase
- sports: frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket
- kitchen: bottle, wine glass, cup, fork, knife, spoon, bowl
- food: banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake
- furniture: chair, couch, potted plant, bed, dining table, toilet
- electronic: tv, laptop, mouse, remote, keyboard, cell phone
- appliance: microwave, oven, toaster, sink, refrigerator
- indoor: book, clock, vase, scissors, teddy bear, hair drier, toothbrush

B Additional Information on Linguistic Feature Extraction

The example below shows the desired tagging and grammatical role parsing (obtained through spaCy) of part of a caption, with the entity mention in color, and the other nouns of the caption in bold.

	A	**boat**	glides	across	the	**water**
Noun order		1				2
Gram role		subj				pobj

We use special heuristics to refine grammatical role. If a noun is labeled as a conjunct (e.g. "*A cat[nsubj] and a dog[conj] are sleeping.*"), we consider the grammatical role label of *dog* to be that of its nearest non-conj parent, in this case *cat[nsubj]*. Similarly, compounds (e.g. "*A girl looks at the television[compound] screen[dobj]*") are labeled with the grammatical role label of their nearest non-compound parent.

As for identifying the noun mentioning the target entity (if any), after a first manual inspection of the automated model identification, we observed that the two main common errors were mapping "skateboarder" to the object "skateboard", and not identifying Golden Retrievers as dogs. We added an automatic filter for those cases.

C Statistical models

Our full model is implemented in R using the `lme4` package for mixed-effects models, `lmer` for linear and `glmer` for binomial models.

```
outcome ~ log_size*source
        + category*source
            + (1 | image)
```

where `source` can be human or one of the three VLMs. The predicted outcome can take the numerical variable `rank_norm` (LMM, Gaussian), and the binary variables `is_mentioned` and `is_subject` (GLMM, Bernoulli/logit).
`rank_norm` corresponds to the noun index divided by the sentence length. `rank_norm` and `is_subject` are conditioned over the noun being mentioned.

The ablation models are computed as follows:
Without category:

```
outcome ~ log_size*source
            + (1 | image)
```

Without size:

```
outcome ~ category*source
            + (1 | image)
```

Without source interactions:

```
outcome ~ log_size
        + category
           +source
      + (1 | image)
```

The coefficients for the full model are visualized in Figure 5, and Table 3 reports which of those are significant.

D Picture intervention prompts

This section presents the different prompts provided to Qwen to perform the intervention on COCO pictures.

Perceptual salience prompts:

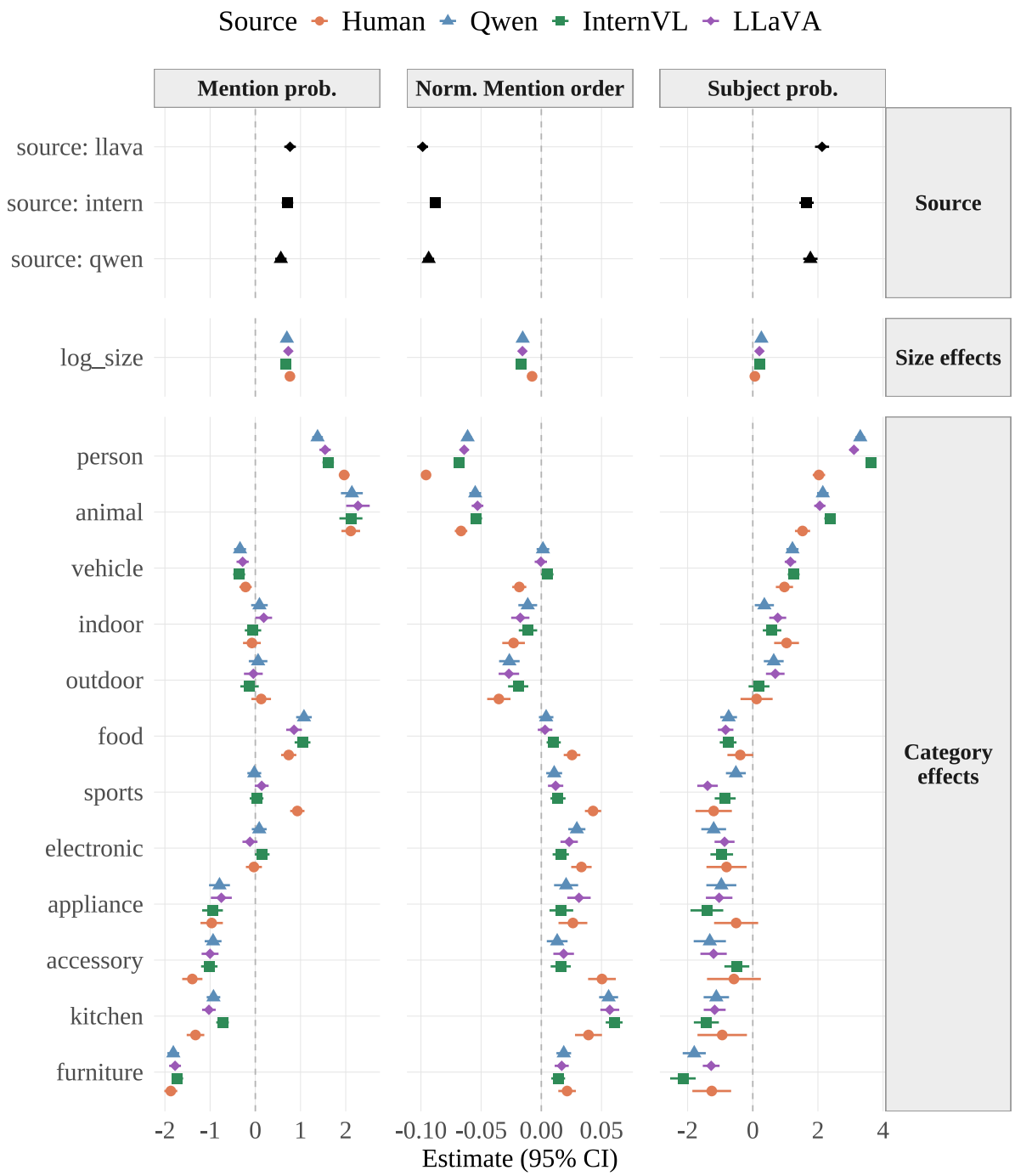


Figure 5: Effects of source, size and category in mixed-effects models fitted to predict mention probability, mention order, and subject probability of entities. The category effects are plotted against the source-specific grand means to facilitate comparison between sources, and show that although the differences between AI and human for each category are often significant (roughly, non-overlapping error bars), these source-interaction effects are typically much smaller than the main effect of that category.

Mention Prob.	**Human**	**Qwen**	**Intern**	**LLaVA**
source	1.762 (0.052)***	0.564 (0.064)***	0.712 (0.064)***	0.768 (0.065)***
log size	0.765 (0.016)***	-0.067 (0.020)***	-0.098 (0.019)***	-0.035 (0.020)·
cat:accessory	-1.394 (0.114)***	0.460 (0.140)**	0.376 (0.139)**	0.393 (0.140)**
cat:animal	2.111 (0.105)***	0.024 (0.153)	0.003 (0.159)	0.159 (0.161)
cat:appliance	-0.965 (0.126)***	0.173 (0.160)	0.018 (0.159)	0.214 (0.160)
cat:electronic	-0.031 (0.090)	0.118 (0.113)	0.184 (0.112)	-0.087 (0.114)
cat:food	0.739 (0.086)***	0.337 (0.115)**	0.305 (0.116)**	0.117 (0.115)
cat:indoor	-0.075 (0.100)	0.167 (0.129)	0.024 (0.129)	0.262 (0.130)*
cat:kitchen	-1.323 (0.099)***	0.398 (0.118)***	0.599 (0.115)***	0.297 (0.119)*
cat:outdoor	0.130 (0.111)	-0.068 (0.143)	-0.257 (0.142)·	-0.173 (0.144)
cat:person	1.965 (0.061)***	-0.588 (0.082)***	-0.353 (0.086)***	-0.422 (0.085)***
cat:sports	0.929 (0.081)***	-0.952 (0.101)***	-0.901 (0.101)***	-0.788 (0.102)***
cat:vehicle	-0.215 (0.070)**	-0.123 (0.088)	-0.142 (0.088)	-0.065 (0.089)
Mention Order	**Human**	**Qwen**	**Intern**	**LLaVA**
source	0.203 (0.002)***	-0.094 (0.002)***	-0.088 (0.002)***	-0.099 (0.002)***
log size	-0.008 (0.001)***	-0.008 (0.001)***	-0.009 (0.001)***	-0.008 (0.001)***
cat:accessory	0.050 (0.006)***	-0.037 (0.007)***	-0.034 (0.007)***	-0.032 (0.007)***
cat:animal	-0.067 (0.003)***	0.012 (0.003)***	0.013 (0.003)***	0.014 (0.003)***
cat:appliance	0.026 (0.006)***	-0.006 (0.008)	-0.010 (0.007)	0.005 (0.007)
cat:electronic	0.033 (0.004)***	-0.004 (0.005)	-0.017 (0.005)**	-0.010 (0.005)·
cat:food	0.026 (0.004)***	-0.022 (0.004)***	-0.015 (0.004)***	-0.022 (0.004)***
cat:indoor	-0.023 (0.005)***	0.012 (0.006)*	0.012 (0.006)*	0.006 (0.006)
cat:kitchen	0.039 (0.006)***	0.017 (0.007)*	0.021 (0.007)**	0.018 (0.007)**
cat:outdoor	-0.035 (0.005)***	0.009 (0.006)	0.016 (0.006)**	0.008 (0.006)
cat:person	-0.096 (0.002)***	0.035 (0.003)***	0.027 (0.003)***	0.032 (0.003)***
cat:sports	0.043 (0.003)***	-0.032 (0.005)***	-0.029 (0.005)***	-0.031 (0.005)***
cat:vehicle	-0.018 (0.003)***	0.020 (0.004)***	0.023 (0.004)***	0.018 (0.004)***
Subject Proba.	**Human**	**Qwen**	**Intern**	**LLaVA**
source	-2.766 (0.098)***	1.769 (0.112)***	1.651 (0.112)***	2.127 (0.110)***
log size	0.063 (0.029)*	0.200 (0.035)***	0.160 (0.035)***	0.140 (0.035)***
cat:accessory	-0.579 (0.422)	-0.741 (0.479)	0.089 (0.454)	-0.625 (0.458)
cat:animal	1.527 (0.119)***	0.623 (0.139)***	0.845 (0.139)***	0.531 (0.139)***
cat:appliance	-0.512 (0.345)	-0.454 (0.397)	-0.898 (0.409)*	-0.521 (0.385)
cat:electronic	-0.807 (0.314)*	-0.394 (0.356)	-0.147 (0.348)	-0.060 (0.339)
cat:food	-0.384 (0.201)·	-0.357 (0.229)	-0.373 (0.229)	-0.450 (0.224)*
cat:indoor	1.036 (0.195)***	-0.681 (0.228)**	-0.445 (0.225)*	-0.271 (0.220)
cat:kitchen	-0.939 (0.387)*	-0.181 (0.426)	-0.487 (0.426)	-0.231 (0.415)
cat:outdoor	0.116 (0.252)	0.527 (0.280)·	0.075 (0.284)	0.574 (0.277)*
cat:person	2.031 (0.097)***	1.269 (0.118)***	1.589 (0.119)***	1.075 (0.117)***
cat:sports	-1.201 (0.285)***	0.682 (0.317)*	0.354 (0.322)	-0.188 (0.320)
cat:vehicle	0.972 (0.135)***	0.246 (0.153)	0.284 (0.152)·	0.184 (0.151)

Table 3: Fixed effects on three linguistic outcomes: Mention probability (binomial logit), Order of mention (normalized rank, linear mixed model), and Subject role probability (binomial logit). For each model, the source terms report the intercepts for Human, and the main effect of each source relative to the Human baseline. The log size terms report the baseline coefficient for Human captions and interaction offsets for model-generated sources (Qwen, Intern, LLaVA). The cat: terms list object category interaction effects per source, and the human baseline. Standard errors are indicated in parentheses. Significance codes: $^{\cdot}p<0.1$, $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$.

(a) Disrupts the scene: horse without carriage

(b) Unexpected location of the bench

(c) Unrealistic size of the horse

(d) Incomplete bike

(e) Duplication of the target entity (kite)

(f) Unrealistic location of the donut and duplication

Figure 6: Examples of pictures generated by Qwen but manually discarded because removing the target entity is disrupting the scene (a), added in unexpected or unrealistic locations (b,c), is unrealistic or incomplete (c,d), or is duplicated (e,f).

Perceptual salience 1

Remove the <target-entity>

Perceptual salience 2

Add a <target-entity>

Perceptual salience 3

Add a <target-entity>. It has to be small or in the background.

Relational salience prompts:

Relational salience 1

Add a person standing on the side.

Relational salience 2

Add a person interacting with the <target-entity>.

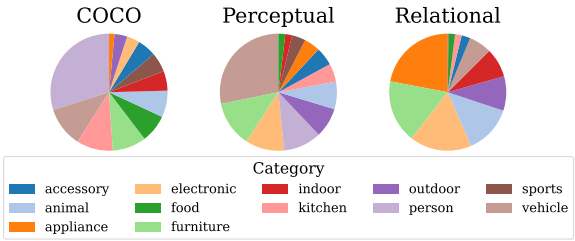


Figure 7: Initial entity distribution in COCO (left), target entity distribution in the perceptual salience dataset (middle) and relational salience dataset (right). Furniture and electronics dominate in both datasets, along with vehicles in the perceptual dataset and appliances in the relational dataset. The distribution shift is due to constraints on image selection, target entity selection criteria, as well as the fact that certain entity modifications were more challenging for the image-editing model.

E Word to category mapping

We provided Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) with the following prompt to identify if any noun of the caption corresponds to the target entity in the picture:

Word to category mapping

Answer with the head of the noun phrase (or the first word in multi-word nouns) mentioning a certain category of entities in a sentence. Only indicate entities that taxonomically belong to a category, not entities that are syntagmatically related. Do not generate anything else. For instance, for the sentence "A Fox Terrier chased the ball across the yard", the noun list ["Fox", "Terrier", "ball", "yard"] and the category "dog", you should generate "Fox". For the sentence. "A skateboarder is captured mid-air performing a trick on a ramp" the noun list ["skateboarder", "air", "trick", "ramp"] you should generate none. Now do the same with the sentence: <sent>, the noun list <[noun list]> and the category <categ>.

F Picture captioning

We used the following models to caption our pictures: LLaVA-OneVision-1.5-8B-Instruct, InternVL3-1B and Qwen2.5-VL-7B-Instruct. We also considered DeepSeek-VL-1.3B-Chat, but could nott get it to generate single-sentence descriptions. The max token generation was set at 128, which created precise descriptions in one sentence only. We provided the model with the following prompts:

Picture captioning

Describe this picture in detail in one sentence. The sentence shouldn't start with: In this picture...

We also tried to prompt the models with the same guidelines as human guidelines provided to COCO annotators, but the models were producing mostly very short captions.

G Examples of discarded pictures and resulting entity distributions

Figure 6 shows examples of pictures manually discarded after picture interventions. It is possible to automatically discard duplications, but incoherence in size and location is more challenging to filter out without human annotations.

Image selection and picture filtering change the entity distribution in both our datasets compared to the initial distribution in COCO. Figure 7 shows the proportion of each entity in our two datasets.

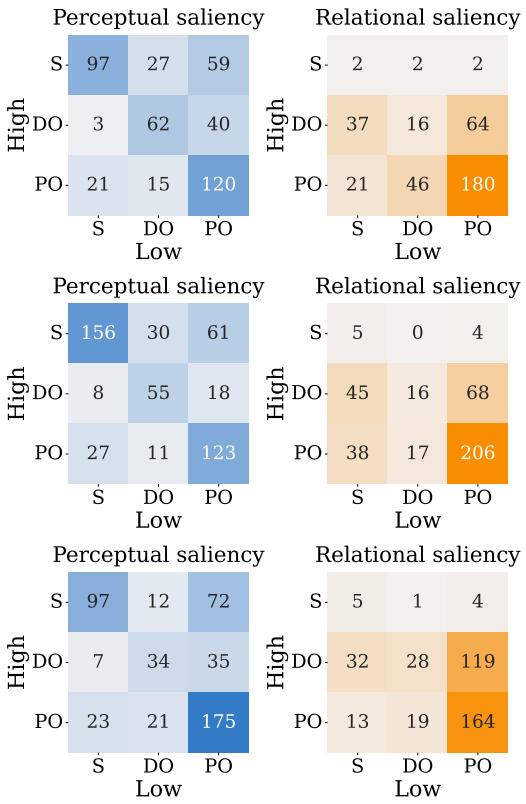


Figure 8: For Qwen (top), LLaVa (middle), and Intern (bottom), the heatmaps represent the counts of the grammatical roles (subject, direct object, or object of a preposition) occupied by target entities in image pairs from the perceptual (left) and relational (right) datasets. Rows and columns in each heatmap indicate the roles occupied when the entity has high or low salience, respectively.

H Causal study results by model

Figure 8 shows the shifts in grammatical role due to changes in salience for each VLM in our study.

Figure 9 shows the mention probability, mention order and grammatical role shift mean values for each of the datasets and the three VLMs.

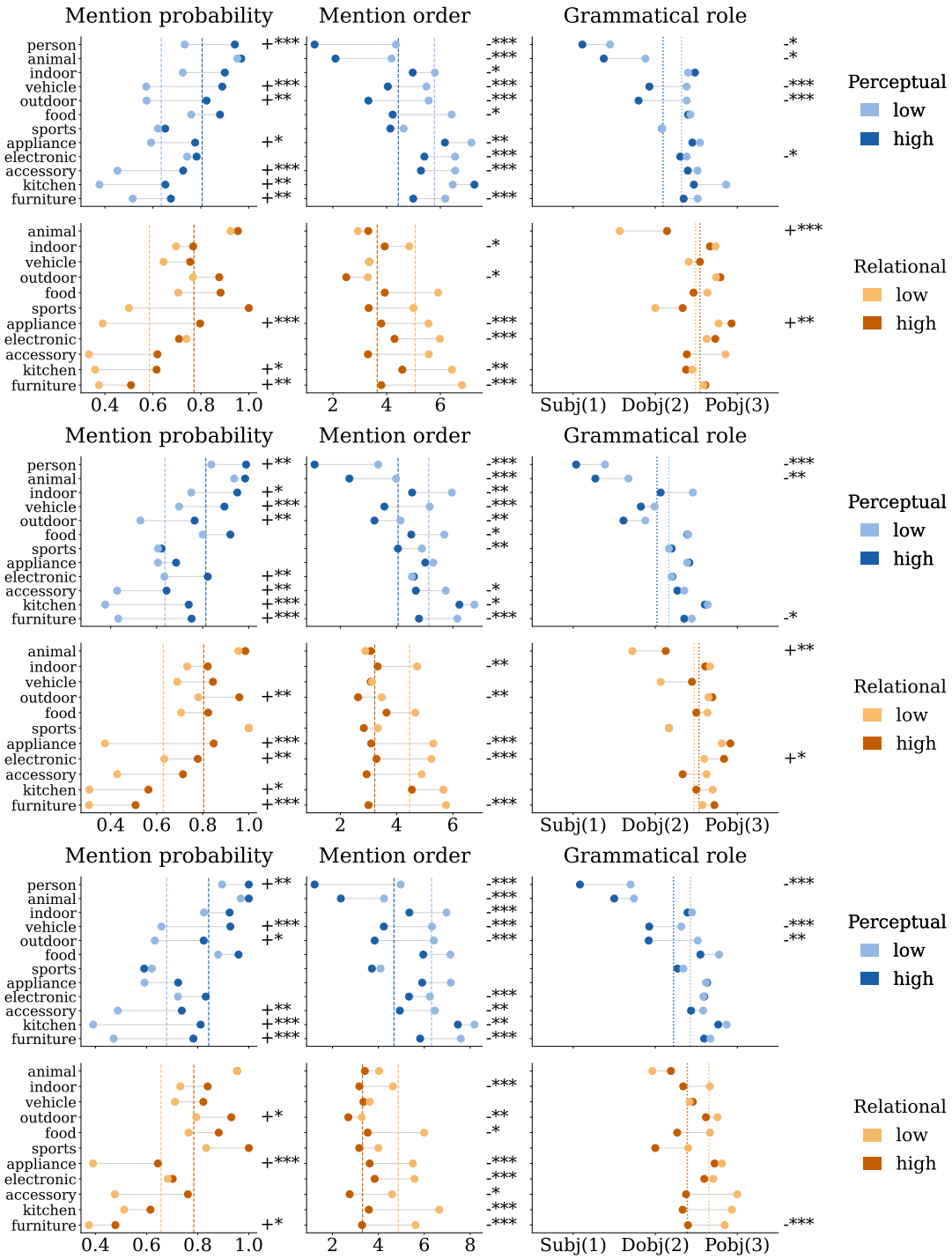


Figure 9: Each double line shows the effects of perceptual salience (top row, blue) and relational salience (bottom row, brown) on mention probability, mention order, and grammatical role, for each of the AI models: Qwen (top), LLaVA (middle) and Intern (bottom). Dots show results for each category, and dotted lines show averages across all categories. To the right of each plot we show significance values (chi-square test for Mention probability and Wilcoxon signed-rank for Mention order and Grammatical role: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$), and +/- to show the direction of significant effects. Grammatical roles are treated as numerical values (1, 2, 3) for visualization, but as ordinal variables in significance tests.