Common Objects Out of Context (COOCo): Investigating Multimodal Context and Semantic Scene Violations in Referential Communication

Filippo Merlo[1,2], Ece Takmaz[1], Wenkai Chen[1], Albert Gatt[1]
[1]Utrecht University [2]University of Trento
f.merlo.research@gmail.com, e.k.takmaz@uu.nl
w.chen5@students.uu.nl, a.gatt@uu.nl

Abstract

To what degree and under what conditions do VLMs rely on scene context when generating references to objects? To address this question, we introduce the *Common Objects Out-of-Context (COOCo)* dataset and conduct experiments on several VLMs under different degrees of scene–object congruency and noise. We find that models leverage scene context adaptively, depending on scene-object semantic relatedness and noise level. Based on these consistent trends across models, we turn to the question of how VLM attention patterns change as a function of target-scene semantic fit, and to what degree these patterns are predictive of categorisation accuracy. We find that successful object categorisation is associated with increased mid-layer attention to the target. We also find a non-monotonic dependency on semantic fit, with attention dropping at moderate fit and increasing for both low and high fit. These results suggest that VLMs dynamically balance local and contextual information for reference generation.

Dataset and code are available here: https://github.com/cs-nlp-uu/scenereg.

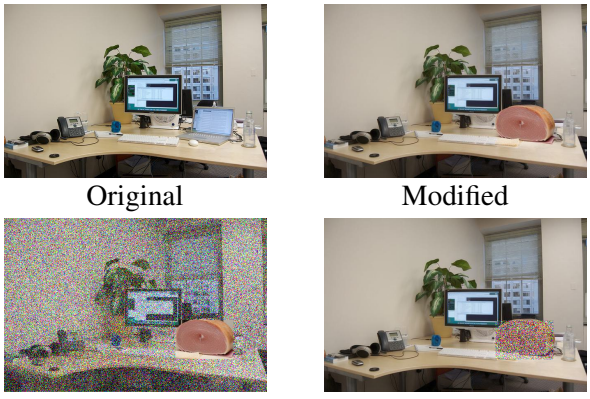
1 Introduction

Objects rarely appear in isolation. Over time, humans build expectations of seeing certain objects in certain contexts. For instance, we expect to see a laptop on a desk in an office, but not a large piece of ham (Figure 1). Studies of human scene perception (Võ, 2021; Torralba et al., 2006; Coco et al., 2016) suggest that scenes exhibit both semantic and syntactic regularities (Biederman et al., 1982; Võ, 2021). Violations of either or both of these result in processing difficulties (Ganis and Kutas, 2003; Öhlschläger and Võ, 2016; Võ, 2021).

In this work, we are interested in how Vision-Language Models (VLMs) behave in the presence of *semantic* violations in visual scenes when referring to objects. We focus on **object naming**, that is, the task of determining the category of an object, a fundamental part of Referring Expression Generation (REG; Krahmer and Van Deemter, 2012). The literature on scene perception suggests that scene semantics mediates human viewers' expectations about the objects in a scene, and viewers deploy their attentional resources accordingly. Scene semantics also influence how humans *describe* objects in scenes (Hwang et al., 2011; de Groot et al., 2017, 2015). Do VLMs similarly utilise scene-level visual information when generating descriptions of objects? To study this, we compare the naming of objects in semantically (in)congruent scenes. If scene-level information guides model choices, then there should be systematic differences between these two conditions in the way models refer to target referents. While previous work has shown that visual context has a facilitating effect in REG (Junker and Zarrieß, 2024), it does not control for the impact of semantic (in)congruity, and relies on small-scale, purposely-trained models or on artificial images (Junker and Zarrieß, 2025). To address these issues, we develop a new dataset, Common Objects Out-of-Context (**COOCo**), which manipulates natural images to introduce objects of varying degrees of consistency with the scene (Figure 1). Similarly to Junker and Zarrieß (2024), we further manipulate the visibility of the target referent and the context by introducing different degrees of noise. Furthermore, since different objects can be related to a scene to different degrees, we model object-scene relatedness as a graded phenomenon. We summarise the contributions of our study as follows.

1. We present COOCo, a novel dataset to evaluate multimodal models' ability to integrate object-level and scene-level visual information, under both congruent and incongruent conditions.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1163–1185, 2026. https://doi.org/10.1162/TACL.a.701
Action Editor: Frank Keller. Submission batch: 6/2025; Revision batch: 12/2025; Published 6/2026.



Original | Modified | Context Noise | Target Noise

Figure 1: An original image of an office scene from the COCO dataset and a *modified* version from COOCo, replacing the laptop with a semantically inconsistent object. Bottom: versions of the modified image with Gaussian noise mask (level = 0.5) on the context and on the target.

2. We investigate the impact of scene context on object identification on several state-of-the-art VLMs. We show that under conditions where target objects violate scene semantics, scene context acts as a distractor, while it has a clear facilitating effect when the target is congruent with the scene. This also makes models resilient to noise when it is applied to the target region.

3. Having established that object-scene relatedness exerts an effect across models, we then focus on a single, representative VLM. Using COOCo and another, recently introduced dataset, we investigate how it deploys attention across scene components under different conditions.

2 Related Work

Context in Referring Expression Generation Referring Expression Generation (REG) is the task of generating identifying descriptions for a target object. Early REG approaches relying on handcrafted algorithms (Krahmer and Van Deemter, 2012) incorporated contextual influences by considering the properties that distinguish a target referent from its distractors, with some research also incorporating constraints to guide coherent categorisation for references to plurals, and the use of landmarks for spatial expressions (Schüz et al., 2023). More recent neural models for *visual REG* learn mappings from raw perceptual inputs to linguistic expressions. Here too there has been limited attention on the ability of models to leverage global scene information (Schüz et al., 2023). This gap has motivated recent investigations into how scene information may enhance robustness in challenging settings (Schüz and Zarrieß, 2023; Junker and Zarrieß, 2025). In particular, Junker and Zarrieß (2024) show that models rely on scene context to generate references even when the target object is fully occluded, indicating that scene context serves both a supportive and disambiguating role during reference generation. Junker and Zarrieß (2025) further find that scene context influences how VLMs describe ambiguous tangram shapes, though to a lesser extent than humans.

Context in Human Object Recognition Human object recognition is strongly influenced by contextual expectations (Bar, 2004; Oliva and Torralba, 2007; Võ, 2021; Peelen et al., 2023), including both global scene properties (e.g., gist, layout) and local cues (e.g., anchor objects, co-occurring items) (Lauer et al., 2018, 2021; Boettcher et al., 2018). Evidence that human overt attention in visual search is semantically constrained comes from viewing tasks, where viewers fixate more on semantically informative (but not necessarily visually salient) regions, and semantic similarity is a strong predictor of gaze trajectories (Henderson et al., 2019; Hayes and Henderson, 2021). Both highly expected and highly incongruent objects attract more attention than moderately fitting ones (Damiano et al., 2024). Motivated by this work, we take a graded view of object-scene semantic similarity in the design of our dataset.

Datasets Widely used vision and language datasets for referring expression and image caption generation, such as COCO (Lin et al., 2015), RefCOCO (Kazemzadeh et al., 2014; Yu et al., 2016) and VisualGenome (Krishna et al., 2016), lack annotations for object-scene relatedness. Scene violation datasets created within the vision community include ObjAct (Shir et al., 2021), Event Task (Dresang et al., 2019), Out-of-Context (Bomatter et al., 2021), Cut-and-paste dataset (Yun et al., 2021), and SCEGRAM (Öhlschläger and Võ, 2016). However, these datasets tend to be relatively small-scale and do not incorporate degrees of semantic violation or occlusion. In this work, we present a new, large-scale dataset for semantic violations in visual scenes, building on existing resources such as COCO-Search18

(Chen et al., 2021), THINGSPlus (Hebart et al., 2019; Stoinski et al., 2023) and SUN (Xiao et al., 2010). Most related to our work are two recent and concurrently released datasets: VISIONS (Allegretti et al., 2025, further described in § 3) incorporates semantic and spatial violations but remains small in scale (1,136 images), providing a single incongruent manipulation per scene. COinCO (Yang et al., 2025) introduces one inpainted object replacement per image and adopts a binary notion of contextual fit based on judgements generated with Molmo (Deitke et al., 2024). In contrast, COOCo offers multiple controlled variants of each scene by systematically modulating object–scene relatedness on a continuous scale, enabling finer semantic manipulation than prior datasets. We also evaluate on VISIONS because it is fully human-validated, provides a continuous object-scene relatedness measure, and includes modal object names.

3 Dataset Creation

We introduce the **Common Objects Out-of-Context (COOCo)** dataset. The dataset provides a collection of images where the semantic coherence of a scene is modified by the presence of a target object with low, medium or high semantic relatedness to the scene type. COOCo builds upon a framework previously established in the visual perception literature with the SCEGRAM dataset (Öhlschläger and Võ, 2017), scaling it up to meet the needs of NLP and computer vision research through a novel methodology.

Images We begin with the publicly available subset of COCO-Search18 (Chen et al., 2021), which includes bounding boxes for target objects. This dataset excludes images with people or animals, categories which are known to significantly influence human visual attention (Clarke et al., 2013; End and Gamer, 2017). Excluding such cues, therefore, avoids confounding factors and enables us to study object-scene semantic relationships. COCO-Search18 further excludes images with targets with multiple object instances, controlling for target size, location, and image aspect ratio. We retained the 2,241 images where the target object is present.

Scene labels For each image, we predict the scene label using a Vision Transformer (ViT)[1] (Dosovitskiy et al., 2021) finetuned on SUN-397 (Xiao et al., 2010). We then reduce the SUN label set (397 labels) to 29 (listed in Appendix A.1), selected to balance frequency and semantic diversity and reclassified the images with the same model, excluding all but the selected labels. This approach mitigates risks of category under-representation and excessive variance. This step provided a scene label for each image.

Object-scene congruency We developed an inpainting pipeline to create versions of images where the target object is replaced with alternatives of varying congruence with the scene (low and medium). We also retain the original image, and additionally include controls where (a) the target is replaced with an object that has high relatedness to the scene, and (b) the target is replaced with a new object of the same category as the original. The latter are included to verify if the effects found in the experiments are due to artefacts caused by inpainting in general, or due to the introduction of incongruent object types. Figure 2 shows an example of a scene and its variants in COOCo.

New objects are determined using human-generated norms in THINGSplus (Stoinski et al., 2023), which includes typicality ratings (the degree of representativeness of an object for higher-level categories) and ratings of the real-world size of objects. We retain only typical objects (typicality ranging from 0.3 to 1) and remove animate objects to ensure consistency with COCO-Search18. We then match COCO-Search18 targets to candidate objects in THINGSPlus with similar real-world size norms, excluding any candidates differing in size from the targets by more than 25 units. We compute semantic relatedness between candidate replacements and scene labels following Hayes and Henderson (2021), who found strong correlations between embeddings in ConceptNet Numberbatch (Speer et al., 2018) and human visual attention shifts. We use the cosine similarity between scene and object label embeddings to filter low-, medium- and high-similarity candidates.

Image modification Given an image, a scene label and a target object with its bounding region and segmentation-based mask obtained from COCO-Search18, we remove the target from the image, and generate 15 new versions of the scene where the target is replaced with objects having

[1] https://huggingface.co/google/vit-base-patch16-224

Figure 2: A set of *COOCo* images from the "cubicle office" scene category with target object *laptop*. The target is removed from the original image ('clean') and replaced with objects of the same type ('generated'), as well as targets with high-, low- and medium relatedness to the scene.

low, medium or high similarity to the scene label (see Figure 2). We also generate an image containing the target as in the original image as a control condition. Only images that pass a verification test are retained. Full details of the inpainting pipeline and verification tests are in Appendix A.1.

Final COOCo dataset The final COOCo dataset comprises 18,395 images: 1,862 original images with corresponding object-removed versions; 5,480 generated images with low relatedness; 5,572 with medium relatedness; 1,771 with high relatedness; and 1,848 images for the same-target condition.

VISIONS dataset We conduct a subset of our experiments on the VISIONS dataset (Allegretti et al., 2025) and compare the results obtained on COOCo with those on it. VISIONS comprises 1,136 naturalistic indoor images constructed from 165 scene setups across seven scene categories: bathroom, bedroom, church, kitchen, office, restaurant, and waiting room. Each setup is photographed in four principal versions created by manipulating the semantic consistency (consistent vs. inconsistent) of a target object, identified with bounding box coordinates, and its spatial position (left vs. right).[2] VISIONS provides an extensive set of perceptual and conceptual norms for both objects and scenes, including object-consistency ratings, which quantify how contextually appropriate the critical object is in each scene. We utilise these ratings. VISIONS also provides human-annotated target objects with free-text labels; we use the modal name for each object as the reference label in our analysis, together with the original object label defined by the dataset creators.

[2] Although the dataset also includes additional swap-based variants for 119 scenes where the target object was exchanged with another object in the scene to produce four mirrored counterparts of the canonical versions, we use only these four canonical versions, as swap manipulations are not part of our analyses.

4 Experimental Setup

We design our experiments to address how VLMs leverage contextual information for object naming, an essential component of Referring Expression Generation. (1) First, we evaluate to what extent VLMs rely on scene context under conditions of varying degrees of semantic congruence between an object and its background; and on how robustly they perform REG under different levels and configurations of visual noise. These experiments are conducted on a variety of SOTA VLMs on the COOCo dataset. Each model is provided with an image and a specified target region, and is prompted to name the object within that area using a structured text prompt. Anticipating the outcomes of this analysis (see § 5), we identify a number of trends across models, showing that they are sensitive to scene semantics when referring to targets and that scene context supports REG when target objects are obscured by noise (cf. Junker and Zarrieß, 2024). (2) Having established these trends across multiple models, we select a single

performant model for in-depth analysis of its outputs (§ 6.1) and its attention deployment across different experimental settings (§ 6.2). This analysis is performed on our own dataset, COOCo, with replications on the VISIONS dataset (Allegretti et al., 2025). This allows us both to generalise findings and control for possible artefacts introduced by inpainting in COOCo.

4.1 Models

We use the following models in our experiments:

KOSMOS-2 (Peng et al., 2024) (1.6B) incorporates mechanisms to establish direct associations between textual spans and specific image regions. It is trained on the GrIT (Grounded Image-Text pairs) web-scale dataset.

Molmo (Deitke et al., 2024) (7B) is trained on a dataset including ca. 2.3M human-annotated pairs of questions and references to image regions based on 2D points, from 428k images.

xGen-MM-Phi3/BLIP-3 (Xue et al., 2024) (ca. 4.4B) is trained on the BLIP3-GROUNDING-50M dataset and supports spatial encoding using bounding box coordinates, descriptive spatial relations and/or relative positioning.

LLaVA-OneVision (Li et al., 2025) (0.5B and 7B) was trained on data with bounding boxes, incorporating RefCOCO (Kazemzadeh et al., 2014; Yu et al., 2016)[>50k samples] and Visual Genome (Krishna et al., 2016)[86k] for single-image settings, and multi-image datasets including WebQA (Chang et al., 2022)[9.3k]. The model is instruction-tuned to accept spatial information directly in text input in the form of normalised bounding box coordinates.

Qwen2.5-VL (Bai et al., 2025) (7B) is trained with coordinates based on images' native resolution for grounding, detection, pointing and object counting. The model accepts absolute coordinates for image regions as part of its prompt.

For detailed descriptions of preprocessing steps and region-of-interest (ROI) prompt formats for each model, refer to Appendix A.2.

Prompts We test models with structured prompts (details in Appendix A.2) with an explicit reference to a region, for example: *What is the object in this part of the image [x_1, y_1, x_2, y_2]? Answer with the object's name only. No extra text.*

4.2 Noise injection

We add Gaussian noise ($\lambda = 0.0, 0.5, 1.0$), ranging from mild distortion to maximum information loss, to the target region, the context region or the entire image ('All noise'). Pixel values in the affected area are perturbed with noise sampled from a normal distribution and clipped to stay within valid image ranges.

4.3 Metrics

We evaluate models using three metrics:

RefCLIPScore (Hessel et al., 2022) is the harmonic mean between (1) the `CLIPScore` (Hessel et al., 2022), which captures the semantic alignment between the generated caption and the image; and (2) the highest cosine similarity between the CLIP (Radford et al., 2021) embeddings of the generated caption and any reference caption. This quantifies the alignment between a generated description for a target and both its visual features (extracted from its image patch) and correct label.

Text-Based Semantic Similarity is computed as the cosine similarity between CLIP text embeddings. We use this measure to assess the semantic similarity between the model's output and the target label (from which Accuracy is derived), as well as between the model's output and the scene label in § 6.1.

Accuracy, defined as 1 if the cosine similarity between the model's output and the target label exceeds a threshold (0.9), and 0 otherwise. To illustrate, consider the target *car*. While the output *car* scores 1, related terms *bus* (0.86), *truck* (0.89), and *police car* (0.89) score 0. In contrast, variants such as *SUV* (0.90), *a race car* (0.90), and the part–whole expression *car door* (0.90) surpass 0.9 and are counted as correct.[3]

5 Results across models

In this section, we report and discuss the main results across all models and experimental conditions in COOCo.

Models are sensitive to scene semantics when referring to targets. Figure 3 reports RefCLIPScore results for each model at noise level 0 across

[3]We also computed a hard accuracy metric based on Levenshtein similarity, which quantifies the character-level edit distance between generated and reference captions. Since results are very similar to those obtained with the cosine-based metric, we do not report them here.

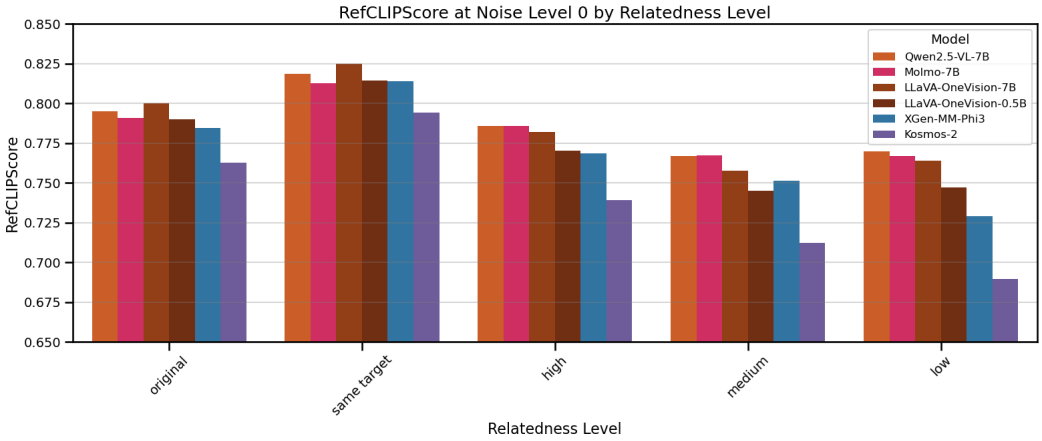
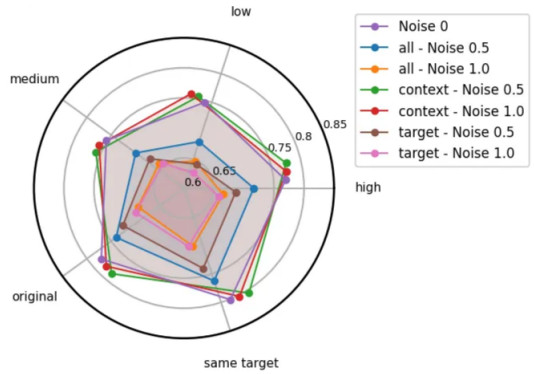
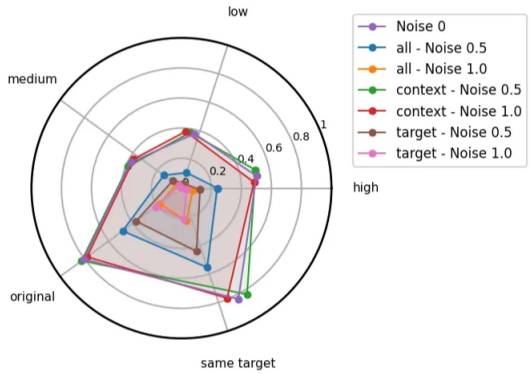


Figure 3: RefCLIPScores per model at noise level 0 across relatedness conditions. Models perform best in the original and same target conditions, with performance declining as target relatedness decreases.



(a) RefCLIPScores by relatedness, noise area, and noise level.



(b) Accuracy across noise levels, noise areas, and relatedness conditions.

Figure 4: RefCLIPScore and Accuracy across experimental conditions, aggregated over all models. When differences within a condition are minimal, the leftmost point corresponds to the higher value, followed by lower values to the right.

all relatedness conditions. Models obtain the highest scores in the *original* and *same target* conditions, indicating stronger alignment with the reference when the target object is preserved or replaced by an object from the same category. Interestingly, performance is higher in the same target condition than in the original. A possible reason for this is that inpainting introduces subtle artefacts that make it easier for target objects to be picked out. Performance decreases when the target object differs from the original (*high*, *medium*, and *low* relatedness). Within these conditions, models perform best under high relatedness, with scores progressively declining as relatedness decreases. Importantly, the observed effects of object–scene relatedness are also supported by results on VISIONS, where no inpainting is applied (see § 6).

For models depicted in warm colours (the first four in Figure 3), the low relatedness condition yields performance that is similar to, or slightly higher than, the medium condition. Conversely, models in cool colours (XGen-MM-Phi3 and Kosmos-2) show a more consistent decline in performance with decreasing relatedness.

Scene context acts as a distractor for targets that violate scene semantics. Figure 4a reports RefCLIPScore aggregated across models by relatedness, noise area, and noise level. Overall, models achieve their best performance at context noise 0.5, followed by noise 1.0 and then the zero-noise condition, with the notable exception of the low-relatedness condition, where noise 1.0 yields the highest alignment. Consistent trends are observed in the accuracy results (Figure 4b): the *origi-*

nal and *same target* conditions reach the highest accuracy (∼0.8) across noise settings, followed by the high-relatedness condition (∼0.5) and the medium and low-relatedness conditions (∼0.4). While most relatedness conditions peak at context noise 0.5, the low-relatedness condition shows a monotonic improvement with increasing noise, indicating that maximal removal of contextual information is particularly beneficial when targets violate scene semantics and scene context becomes detrimental.

Scene context acts as a facilitator when congruent targets are obscured. In contrast, applying noise directly to the target region results in a clear performance degradation, proportional to the noise intensity: the higher the noise level, the greater the drop. When noise (at 0.5) is applied to the target or the entire image, accuracy in the original and same target conditions drops to 0.5, while all other conditions approach zero (Figure 4b). For RefCLIPScore, we find that at 0.5 noise level, for the *high*, *medium*, and especially *low* relatedness conditions, performance is more robust when noise is applied to the entire image rather than solely to the target. At a higher noise level (1.0), performance continues to decline. However, models perform best with target-specific noise in the *original* and *same target* conditions. This suggests that, given a non-perturbed scene context, model predictions for noised target regions are for objects that would be expected given the scene composition (see § 6.1). Comparing the 0.5 and 1.0 all-noise settings, we see a sharper drop for the original and same-target conditions, while the drop is smaller for the high, medium and especially the low relatedness conditions.

Mixed-Effects Analysis of RefCLIP Similarity Focusing on the 'extreme' semantic conditions (*original* and *low* relatedness), we build a Linear Mixed-Effects (LME) model with RefCLIPScore as the dependent variable, noise level (scaled and centred) as a continuous predictor, and noise area as a categorical factor contrasting context and target noise against the *all-noise* baseline. The model includes random intercepts for the image identifier and the VL models.[4] Table 1 reports the LME results. Confirming our observations, baseline RefCLIPScore is higher for the original than for the low-relatedness images. Context noise increases similarity, while target noise decreases it. As expected, this contrast is stronger in the low-relatedness condition. Noise intensity shows a significant negative main effect, with interactions showing a less steep negative impact when noise is in the context or the target compared to full-image noise.

6 In-depth analysis of LLaVA-OneVision

We follow up with further analysis of LLaVA-OneVision-0.5B, chosen because of its size, which makes in-depth exploration feasible and its competitive performance (cf. Figure 3; full accuracy scores for this model are in Appendix B, Tab 6 and 7). We report analyses on both COOCo and VISIONS (Allegretti et al., 2025), rescaling images in the latter from the original (1920 × 1080 px) to match the resolution in COOCo (max. width of 640 px). We note that for VISIONS, naming differences between dataset labels and human-provided modal object names have negligible impact on accuracy (see Appendix A.3).

6.1 Output vs. scene similarity

We begin with an analysis of the text-based semantic similarity metric scores between the model's predicted reference to the target (e.g. *laptop*), and the scene label for an image (e.g. *office*), to evaluate whether the generation was more strongly anchored to the overall scene or to the target object itself. Figure 5 reports how semantic similarity between the model's outputs and the scene labels varies across correctness, noise conditions and intensity, and the relatedness distinctions defined in each dataset: high vs. low relatedness in COOCo, and congruent vs. incongruent scenes in VISIONS.[5]

When the target region contains *no noise* (in the baseline condition with 0 noise, and in the context-noise setup), the same trend appears in both correct and incorrect outputs: in low-relatedness conditions, the model produces predictions that are less semantically aligned with the scene label than in the higher-relatedness conditions of each dataset. This indicates that, when the target object violates scene semantics, predictions are less anchored to the global scene and are instead more

[4] Since we are interested here in establishing the statistical reliability of the trends observed above across all models.

[5] Appendix B, Fig. 10 presents a complementary analysis, treating object-scene relatedness as continuous rather than categorical.

Predictor	Original images			Low relatedness		
	Coef.	Std.Err.	z	Coef.	Std.Err.	z
Intercept	0.735	0.002	379.21*	0.692	0.003	275.51*
Noise area (context)	0.052	0.001	94.13*	0.062	0.000	141.32*
Noise area (target)	-0.015	0.001	-27.34*	-0.047	0.000	-107.65*
Noise level (cont.)	-0.034	0.000	-145.42*	-0.034	0.000	-182.28*
Noise area (context) × Noise level	0.032	0.001	60.75*	0.036	0.000	86.79*
Noise area (target) × Noise level	0.014	0.001	26.76*	0.024	0.000	56.84*
Group Var	0.041			0.069		

Table 1: Linear mixed effects model: Noise area and noise level on RefCLIPScores, for original and low-relatedness images. * denotes a significant effect at $p \leq .001$.

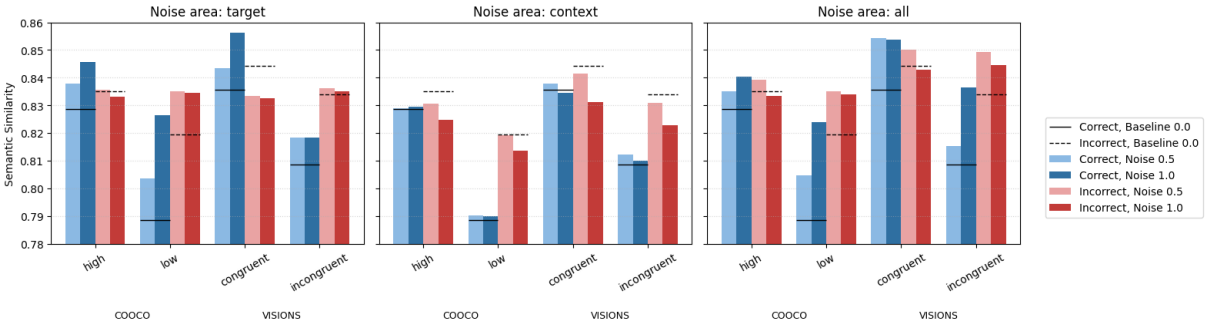


Figure 5: Semantic similarity between the model's outputs and scene labels for LLaVA-OneVision-0.5B in COOCo (ours) and VISIONS (Allegretti et al., 2025), shown across correctness (blue: correct; pink/red: incorrect), semantic-fit conditions (COOCo: high/low; VISIONS: congruent/incongruent), noise areas (target, context, all), and noise levels (0.5, 1.0). Dashed and solid horizontal lines indicate baseline similarity at zero noise for incorrect and correct predictions, respectively.

influenced by the unrelated object itself. Crucially, this pattern disappears when noise is applied to the *target region*. Incorrect predictions under target noise no longer exhibit clear differences between high and low semantic fit, suggesting that once the target's appearance is degraded, object semantics cease to guide outputs. In contrast, correct predictions under target noise consistently shift towards the scene label: across all semantic-fit conditions, correct responses exhibit higher similarity to the scene than in the no-noise baseline (solid line). This aligns with the observation in § 5 that degrading the target prompts the model to compensate by relying more heavily on global scene context. When successful, the resulting predictions are more semantically aligned with the scene. Notably, only in the congruent case of the VISIONS dataset, incorrect predictions fall below their zero-noise baseline, suggesting that congruent objects, when not degraded, may anchor the answers to a higher semantic similarity with the scene.

When noise is applied to the *context region*, correct responses exhibit no meaningful deviation from the zero-noise baseline, indicating that when the target remains intact, the model continues to anchor its prediction primarily to the object itself rather than the broader scene. Incorrect responses, however, show a general decrease in similarity relative to the baseline, suggesting that context degradation harms scene-based anchoring when the model fails to identify the target correctly. Under *all noise* conditions, the effects differ across datasets: COOCo exhibits patterns comparable to those observed under target noise alone, whereas in VISIONS, the incorrect predictions in congruent trials maintain their baseline similarity to the scene, and those in incongruent trials even show an increase in similarity.

We further analyse these results using LME on the COOCo data; the same analysis on VISIONS is reported in Appendix B, Tab 5. We treat semantic similarity (logit-transformed) between predicted target and scene label as the dependent variable. Noise level (modeled as a continuous variable, scaled and centred) and area (comparing target and context to the *all* noise condition) are

	Original images			Low relatedness		
	Coef.	Std.Err.	z	Coef.	Std.Err.	z
Intercept	1.639	0.044	37.538*	1.598	0.034	46.716*
Noise: Context	-0.006	0.007	-0.787	-0.183	0.004	-41.733*
Noise: Target	-0.048	0.007	-7.364*	-0.032	0.004	-8.042*
Noise level (cont.)	-0.068	0.007	-10.377*	-0.004	0.004	-1.062
Noise level × Noise: Context	0.051	0.009	5.472*	-0.016	0.006	-2.867*
Noise level × Noise: Target	0.039	0.007	5.364*	0.066	0.004	14.865*
Group Var	0.047	0.048		0.029	0.029	

Table 2: Linear mixed effects model: Noise area and noise level on scene-target semantic similarity, for original and low relatedness contexts. * denotes a significant effect at $p \leq .01$. Conditional R^2 values: 0.52 (original) and 0.39 (low rel.)

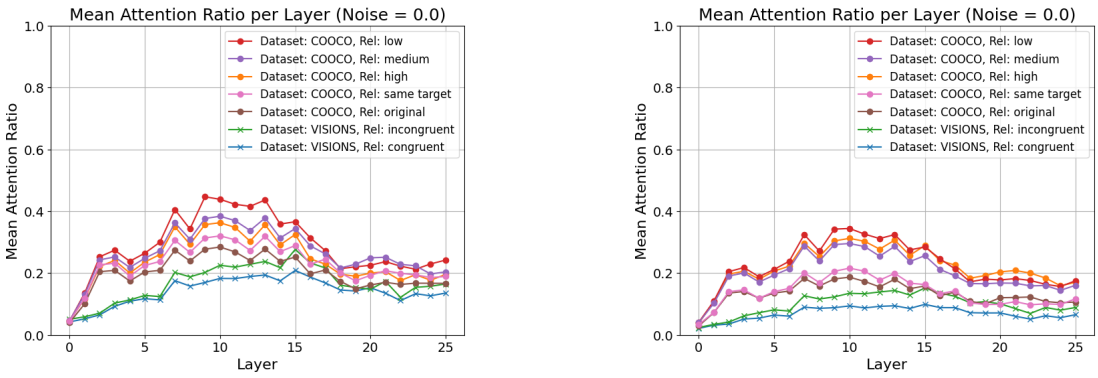
predictors, with random intercepts for scene labels.[6] We include both correct and incorrect outputs, analysing separately the two 'extreme' cases of target-scene congruence, namely, the *original* and the *low-relatedness* conditions. Table 2 summarises the LME results. The main effect of noise level is significant in the *original*, but not the *low relatedness*condition. The *target* noise area is signficantly different from the *all* noise condition, but *context* is only significant in low-relatedness images. (All main effects are significant in VISIONS; cf. Tab. 5). More important are the interactions: In original images, increased noise in the target or context region results in slight but significant increases in semantic similarity of the predicted target to the scene label. In the low relatedness condition, increasing noise in the context region results in a significant decrease in similarity between model predictions and scene label.

6.2 Attention Deployment Analysis

We now turn to an analysis of the model's attention dynamics under different experimental conditions. In what follows, we use the term 'attention' to refer to the Transformer decoder's attention to the image tokens when it generates the name of the object. To analyze attention allocation across different layers l of the vision encoder and image samples t, we compute the ratio r between the total normalized attention values assigned to the target (a_{target}) and the context (a_{context}): $r_l^{(t)} = \frac{a_{\text{target},l}^{(t)}}{a_{\text{context},l}^{(t)}}$. A value of r higher than 1 indicates a higher allocation of attention to the target region, whereas a value closer to 0 suggests greater attention to the context. Since the context area is larger than the target, we do not expect r to exceed 0.5. Instead, our focus is on analysing its variations across different experimental conditions. We then compute the mean ratio value for each encoder layer by averaging across all samples: $\bar{r}_l = \frac{1}{T}\sum_{t=1}^{T} r_l^{(t)}$ where T denotes the total number of instances. This allows us to assess how attention is distributed across different hierarchical levels of the encoder under varying experimental conditions. Full details of how we extract attention values and compute layerwise attention ratios are in Appendix A.4.

Targets that violate scene semantics attract more attention. Figure 6 shows the attention allocation ratio across layers $\bar{r}_l$ for different semantic relatedness conditions at zero noise, separated by correct and incorrect responses according to the accuracy metric (§ 4.3). We note several trends that hold in both COOCo and VISIONS datasets. First, we observe that, for correct (resp. incorrect) responses, the model generally applies more attention to the target. Second, target-focused attention is primarily concentrated in the middle layers, peaking between layers 7 and 15. Third, the model's attention to the target is clearly modulated by its semantic congruency with the scene: more attention is allocated in the *low-relatedness* condition, followed by the *medium-relatedness* condition, and least in the *high-relatedness* condition. The least attention is allocated to the target in the *original* condition, while the *same target* condition receives slightly more. Interestingly, this pattern of increased attention to less semantically related targets also appears in incorrect responses. This suggests that even when the model fails to correctly name the object, it may still detect its incongruity with the surrounding scene, allocat-

[6] We do not include a full random effects structure, as introducing random slopes resulted in poor convergence.



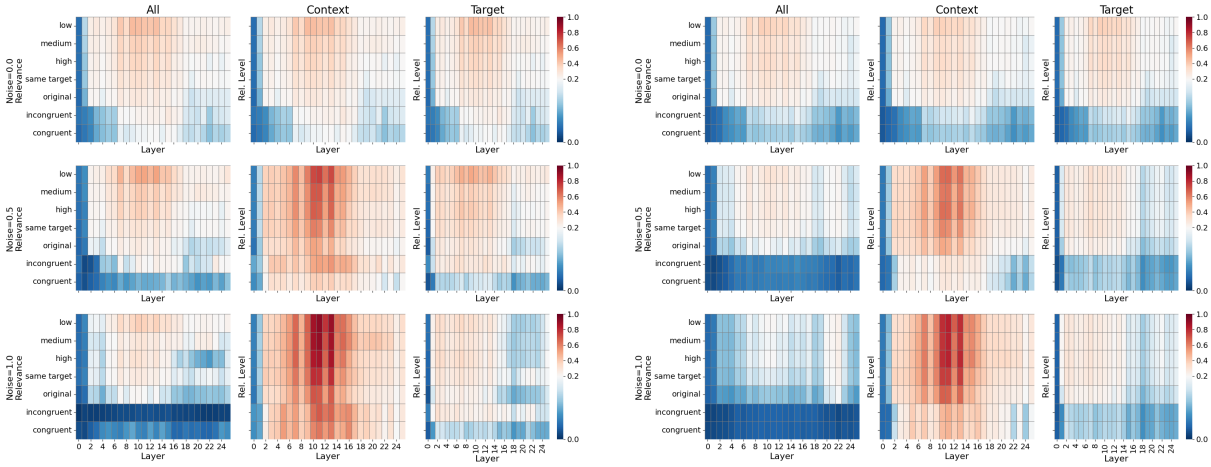
Correct Responses **Incorrect Responses**

Figure 6: Layer-wise attention allocation ratio on the target object across semantic relatedness conditions at zero noise, split by correct and incorrect responses.

ing more attention to targets that appear out of place. While the above observations apply to both datasets, we also note that $\bar{r}_l$ is generally lower in VISIONS. This is unlikely to be due to cues introduced by inpainting, as the difference also holds in comparison to the *original* image condition in COOCo. We hypothesise that the lowered attention-to-target in VISIONS is either an artefact of the image compression performed for our experiments, or is related to general properties of the images (e.g. the relative size of target regions). Deeper comparison of the differences between the two datasets is left to future work.

Targets attract attention even under moderate noise conditions. Figure 7 compares mean attention-to-target ratio $\bar{r}_l$ heatmaps for correct versus incorrect model responses across the three noise levels (0.0, 0.5, 1.0) and three noise conditions (*all, context, target*). The analysis shows that moderate noise on the target (e.g., noise level 0.5 in the *all* and *target* noise conditions) tends to increase attention to the target in correct responses, particularly in the middle layers (7-14), suggesting that the model compensates for partial degradation by focusing more on the target. However, when the noise becomes too severe, attention to the target declines, likely because the target information becomes less recoverable. In contrast, under the *context* noise condition, higher noise levels consistently lead to increased attention on the target, indicating a reallocation of focus away from the noisy context toward the unaltered target. These attention patterns appear consistently in both COOCo and VISIONS. Once again we note a subtle difference between the two datasets: in the *all* noise condition, there is virtually no attention to the target region in the VISIONS samples. One possible reason, already alluded to above, is that inpainting in COOCo creates subtle cues that facilitate attention to the target region. However, in the COOCo original (non-inpainted) images we do observe some attention to the target, albeit much less so than in other conditions. These observations suggest that artefacts due to inpainting may facilitate attention to the target under *all* noise, but they are not the sole determining factor.
Finally, we also see that when the model generates a correct response, there is higher attention to the target region, also under noisy conditions. Visualisations are in Appendix B (Fig. 11).

Sensitivity to scene-target consistency is non-monotonic We further inspect how the attention-to-target ratio ($\bar{r}_l$) varies as a continuous function of scene–object semantic fit. For both COOCo and VISIONS, we restrict analyses to the samples where the model correctly identifies the target. In COOCo, we focus on manipulated images and exclude the *original* and *same-target* controls, as these do not involve modifications to the scene semantics; in VISIONS, we use the human-validated *object-consistency ratings*. We compute the mean attention-to-target ratio for each layer across the selected samples and identify the subset of layers that consistently show higher target focus (top quartile by mean attention). For each image, we then obtain a single attention score by averaging the attention-to-target ratio over this selected subset of layers. Semantic relatedness (COOCo) and object consistency (VISIONS) are normalised to the $[0, 1]$ range and



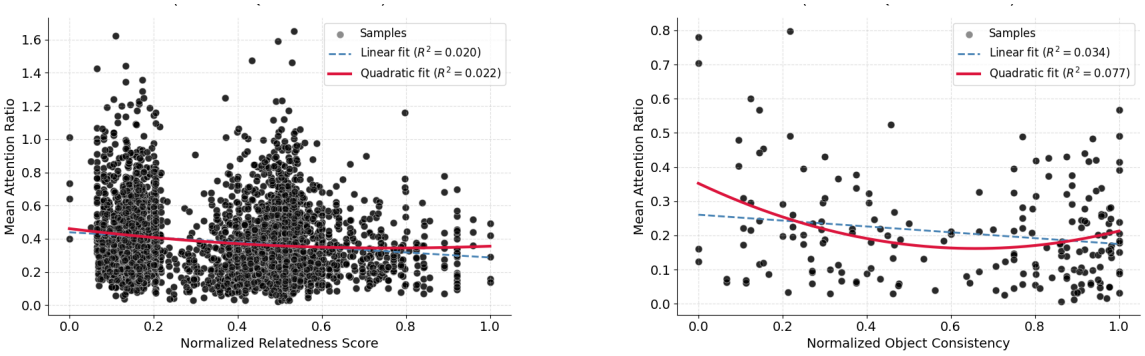
Correct Responses **Incorrect Responses**

Figure 7: Layer-wise attention heatmaps showing how the model's focus on the target varies with noise level (rows: none; moderate with $\lambda = 0.5$; high with $\lambda = 1.0$) and noise location (columns: all, context only, target only). Warmer colors indicate stronger attention relative to the no-noise baseline (average attention ratio ∼0.19), while cooler colors indicate reduced attention to the target. Target-scene relatedness levels range from *low* to *original* for COOCo, and *in/congruent* for VISIONS.

treated as continuous predictors. We model their relationship to the aggregated attention-to-target ratio using both linear and quadratic regression functions, and compare their goodness-of-fit via R^2. We present results for the 0 noise condition in Fig. 8; for results for the noised conditions refer to Appendix B, Fig. 12. At 0 noise, across datasets, attention toward the target exhibits a non-linear dependency on semantic fit. In VISIONS, a linear model revealed a significant negative association between object consistency and attention ($\beta = -0.0858$, $p < .001$), and adding a quadratic term significantly improved fit ($\beta_2 = 0.4395$, $p < .001$; nested model comparison: $F(1, 597) = 27.75$, $p = 1.93 \times 10^{-7}$). The resulting curve was convex, with a minimum at $x = 0.659$ on the normalized scale, indicating reduced attention for moderately consistent objects and increased attention at higher consistency levels. A similar pattern emerged in COOCo. The linear model again showed a significant negative slope ($\beta = -0.1519$, $p < .001$), and the quadratic term provided a significant improvement ($\beta_2 = 0.2003$, $p < .001$; nested model comparison: $F(1, 12051) = 19.86$, $p = 8 \times 10^{-6}$), with the minimum located at $x = 0.764$. While the effect is weaker in COOCo, both datasets exhibit reliable convex trends, suggesting that the semantic relatedness between target object and scene modulates attention in a non-monotonic manner rather than through a purely linear decline. This pattern is also present in other conditions for different noise levels and areas.

7 Discussion

The RefCLIPScore and Accuracy analyses reveal that models consistently exhibit superior performance when either the original target object or a semantically similar one is present. While it is possible that COCO images were included in models' training data, the observation that scores are slightly higher for the *same target* condition than for the *original* condition suggests that models benefit from small semantic shifts within the same category, potentially because the generated object is more prototypical and thus easier to categorise. Alternatively, inpainting introduces artefacts that may make the target object more salient, thereby drawing additional model attention. This could be mitigated in future work by using more advanced, frontier inpainting models. However, the decrease in performance with reduced object–scene relatedness indicates that models are sensitive to target-context semantic compatibility. Noise manipulations provide additional insight: When targets are weakly related to the scene, reducing contextual information improves performance, whereas for semantically congruent targets, preserving scene



COOCo dataset **VISIONS dataset**

Figure 8: Mean attention-to-target ratio as a function of semantic fit in COOCo (left) and VISIONS (right), aggregated over layers with consistently higher target focus. Points represent individual samples; dashed lines show linear regression fits, and solid red lines show quadratic fits.

context is more beneficial when the target region is degraded.

The output versus scene similarity results for LLaVA-OneVision indicate that as semantic relatedness decreases, the model tends to anchor outputs more to the semantics of the target rather than the broader scene. When target objects are corrupted by noise, correct responses shift towards a stronger reliance on scene context, implying an adaptive strategy that leverages global context to compensate for degraded local features, consistent with Junker and Zarrieß (2024).

In the attention analysis, clear differences related to semantic fit emerge in the Transformer decoder's attention allocation. Correct categorisation consistently involves higher attention to the target, particularly in middle layers (7–15), suggesting that critical semantic processing occurs there, as already noted in interpretability studies on the LLaVA model family (Zhang et al., 2025a,b; Golovanevsky et al., 2025). A compelling observation is the inverse relationship between semantic relatedness and attention allocation: less scene-related targets receive more focused attention. Furthermore, noise conditions modulate attention dynamics significantly. Moderate target-specific noise (0.5 level) triggers a compensatory increase in target-focused attention, aiding correct recognition. Conversely, extreme noise (level 1.0) forces models to shift reliance entirely to context. This dynamic attention behaviour highlights an adaptive mechanism within the model, balancing local target analysis against global context integration depending on image clarity and semantic coherence. Finally, LLaVA-One-Vision's non-monotonic sensitivity to scene–object fit has a notable parallel in work by Damiano et al. (2024), whose eye-tracking study shows that both highly expected and unexpected objects receive increased fixations (see also Çelikkol et al., 2023). The resulting U-shaped relationship between semantic similarity and overt attention has an analogous trend in our results: moderately consistent targets attract reduced attention, whereas both low- and high-consistency targets receive higher focus. While the underlying mechanisms clearly differ, this similarity indicates that the model's attention patterns resemble key aspects of human scene processing.

8 Conclusion

This paper presented COOCo, a dataset to explore scene-semantic violations. Consistent with recent work (Junker and Zarrieß, 2024; Junker and Zarrieß, 2025), we show that VLMs rely on scene semantics to refer to objects, but this is modulated by perturbations to these image regions. While these results hold across different models, and are supported by in-depth analysis of LLaVA-OneVision-0.5B, we believe that detailed comparison of different architectures would shed further light on the role of context in VLMs. We replicated our findings on VISIONS (Allegretti et al., 2025), which differs from COOCo in that it does not involve artificial image manipulation.

Our analysis of attention under different conditions contributes to a growing body of interpretability research aimed to localise cross-modal fusion in VLMs (Zhang et al., 2025a,b; Golovanevsky et al., 2025).

Acknowledgements

This research was supported by the Netherlands Science Foundation (NWO) under the AiNed XS Europe 2023 program, grant number NGF.1609.23.020. FM was a visiting student from the University of Trento to Utrecht University during the period when the work was conducted. We thank Sina Zarrieß, Simeon Junker and Hendrik Buschmeier for their valuable comments and helpful discussions. We also thank our four anonymous TACL reviewers and the action editor, whose input helped to improve the paper.

References

Elena Allegretti, Giorgia D'Innocenzo, and Moreno I. Coco. 2025. The Visual Integration of Semantic and Spatial Information of Objects in Naturalistic Scenes (VISIONS) database: attentional, conceptual, and perceptual norms. *Behavior Research Methods*, 57(1):42.

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. ArXiv:2502.13923 [cs].

Moshe Bar. 2004. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629.

Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. 1982. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177.

Sage E. P. Boettcher, Dejan Draschkow, Eric Dienhart, and Melissa L.-H. Võ. 2018. Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision*, 18(13):11.

Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. 2021. When Pigs Fly: Contextual Reasoning in Synthetic and Natural Scenes. ArXiv:2104.02215 [cs].

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. WebQA: Multihop and Multimodal QA. ArXiv:2109.00590 [cs].

Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. 2021. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Scientific Reports*, 11(1):8776.

Alasdair D F Clarke, Moreno I. Coco, and Frank Keller. 2013. The impact of attentional, linguistic, and visual features during object naming. *Frontiers in Psychology*, 4(DEC):1–12.

Moreno I. Coco, Frank Keller, and George L. Malcolm. 2016. Anticipation in real-world scenes: The role of visual context and visual memory. *Cognitive Science*, 40(8):1995–2024.

Claudia Damiano, Maarten Leemans, and Johan Wagemans. 2024. Exploring the Semantic-Inconsistency Effect in Scenes Using a Continuous Measure of Linguistic-Semantic Similarity. *Psychological Science*, 35(6):623–634.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words:

Transformers for Image Recognition at Scale. ArXiv:2010.11929 [cs].

Haley C. Dresang, Michael Walsh Dickey, and Tessa C. Warren. 2019. Semantic memory for objects, actions, and events: A novel test of event-related conceptual semantic knowledge. *Cognitive Neuropsychology*, 36(7-8):313–335.

Albert End and Matthias Gamer. 2017. Preferential processing of social features and their interplay with physical saliency in complex naturalistic scenes. *Frontiers in Psychology*, 8(March):418.

Giorgio Ganis and Marta Kutas. 2003. An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, 16(2):123–144.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. 2025. What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation. 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 11462–11482, Albuquerque, New Mexico. Association for Computational Linguistics.

Floor de Groot, Falk Huettig, and Christian N. L. Olivers. 2015. When Meaning Matters: The Temporal Dynamics of Semantic Influences on Visual Attention. *Journal of Experimental Psychology: Human Perception and Performance*.

Floor de Groot, Falk Huettig, and Christian N.L. Olivers. 2017. Language-induced visual and semantic biases in visual search are subject to task requirements. *Visual Cognition*, 0(0):1–16. Publisher: Taylor & Francis.

Taylor R. Hayes and John M. Henderson. 2021. Looking for semantic similarity: What a vector-space model of semantics can tell us about attention in real-world scenes. *Psychological Science*, 32(8):1262–1270. PMID: 34252325.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792.

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.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. ArXiv:2104.08718.

Alex D. Hwang, Hsueh-Cheng Wang, and Marc Pomplun. 2011. Semantic guidance of eye movements in real-world scenes. *Vision Research*, 51(10):1192–1205. Publisher: Elsevier Ltd.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2017. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. ArXiv:1712.05877 [cs].

Simeon Junker and Sina Zarrieß. 2024. Resilience through scene context in visual referring expression generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 344–357, Tokyo, Japan. Association for Computational Linguistics.

Simeon Junker and Sina Zarrieß. 2025. SceneGram: Conceptualizing and Describing Tangrams in Scene Context. *arXiv preprint arXiv:2506.11631*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Emiel Krahmer and Kees Van Deemter. 2012. Computational Generation of Referring Expressions: A Survey. *Computational Linguistics*, 38(1):173–218.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie

Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. ArXiv:1602.07332 [cs].

Tim Lauer, Tim H. W. Cornelissen, Dejan Draschkow, Verena Willenbockel, and Melissa L.-H. Võ. 2018. The role of scene summary statistics in object recognition. *Scientific Reports*, 8(1):14666.

Tim Lauer, Filipp Schmidt, and Melissa L.-H. Võ. 2021. The role of contextual materials in object recognition. *Scientific Reports*, 11(1):21988.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. ArXiv:1405.0312 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. ArXiv:2304.08485 [cs].

Sabine Öhlschläger and Melissa L.-H. Võ. 2016. Scegram: An image database for semantic and syntactic inconsistencies in scenes. *Behavior Research Methods*, 49:1780–1791.

Aude Oliva and Antonio Torralba. 2007. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527.

Marius V. Peelen, Eva Berlot, and Floris P. De Lange. 2023. Predictive processing of scenes and objects. *Nature Reviews Psychology*, 3(1):13–26.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. ArXiv:2103.00020 [cs].

Simeon Schüz, Albert Gatt, and Sina Zarrieß. 2023. Rethinking symbolic and visual context in Referring Expression Generation. *Frontiers in Artificial Intelligence*, 6:1067125.

Simeon Schüz and Sina Zarrieß. 2023. Keeping an Eye on Context: Attention Allocation over Input Partitions in Referring Expression Generation. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 20–27, Prague, Czech Republic. Association for Computational Linguistics.

Yarden Shir, Naphtali Abudarham, and Liad Mudrik. 2021. You won't believe what this guy is doing with the potato: The ObjAct stimulus-set depicting human actions on congruent and incongruent objects. *Behavior Research Methods*, 53(5):1895–1909.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. ArXiv:1612.03975 [cs].

Laura M. Stoinski, Jonas Perkuhn, and Martin N. Hebart. 2023. THINGSplus: New norms and metadata for the THINGS database of 1854 object concepts and 26,107 natural object images. *Behavior Research Methods*.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. ArXiv:2109.07161 [cs].

Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the Structure of Attention in a Transformer Language Model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Melissa Le-Hoa Võ. 2021. The meaning and structure of scenes. *Vision Research*, 181:10–20.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, San Francisco, CA, USA. IEEE.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. 2024. xgen-mm (blip-3): A family of open large multimodal models.

Tianze Yang, Tyson Jordan, Ninghao Liu, and Jin Sun. 2025. Common Inpainted Objects In-N-Out of Context. ArXiv:2506.00721 [cs].

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling Context in Referring Expressions. In *ECCV*. ArXiv: 1608.00272.

Woo-han Yun, Taewoo Kim, Jaeyeon Lee, Jaehong Kim, and Junmo Kim. 2021. Cut-and-Paste Dataset Generation for Balancing Domain Gaps in Object Instance Detection. *IEEE Access*, 9:14319–14329. ArXiv:1909.11972 [cs].

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. ArXiv:2303.15343 [cs].

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025a. MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs. ArXiv:2502.17422 [cs].

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025b. Cross-modal Information Flow in Multimodal Large Language Models. ArXiv:2411.18620 [cs].

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. 2024. A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting. ArXiv:2312.03594 [cs].

Pelin Çelikkol, Jochen Laubrock, and David Schlangen. 2023. TF-IDF based Scene-Object Relations Correlate With Visual Attention. In *2023 Symposium on Eye Tracking Research and Applications*, pages 1–6, Tubingen Germany. ACM.

Sabine Öhlschläger and Melissa Le-Hoa Võ. 2017. SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. *Behavior Research Methods*, 49(5):1780–1791.

A Additional preprocessing, data and model details

A.1 COOCo image creation

living room, arrival gate outdoor, restaurant, market/outdoor, parking lot, bakery shop, arrival gate/outdoor, art studio, gas station, banquet hall, tower, street, music studio, fastfood restaurant, pantry, cubicle/office, market outdoor, kitchen, coffee shop, bedroom, dorm room, cubicle office, delicatessen, train railway, bakery/shop, home office, physics laboratory, dining room, bathroom.

Table 3: The 25 scene labels in the COOCo dataset

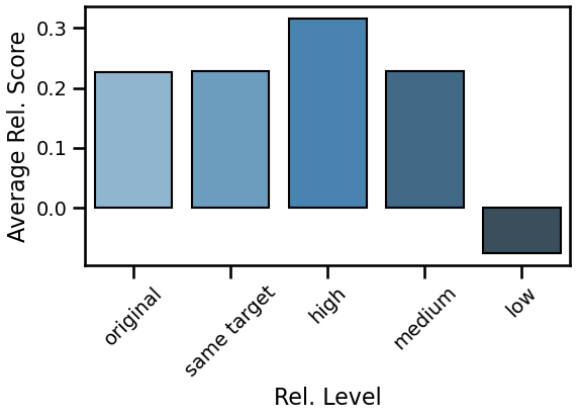


Figure 9: Average relatedness scores grouped by relatedness level, illustrating the semantic similarity between scenes and candidate objects.

A.1.1 Inpainting and verification pipeline

We use the LaMa inpainting model (Suvorov et al., 2021) to remove the target object from a scene, expanding its mask by 20% to make sure that the resulting image is free of irregularities. We then create a square-padded version of the cleaned image, and create a mask for object inpainting, For each of the three relatedness groups (low, medium, and high), we retain 15 replacement candidates. In the high-similarity condition only, the similarity to the original target object is also considered. Additionally, we generate one image containing the same target object as in the original image to use as a control condition. See Figure 9 for the distributon of relatedness scores per relatedness level.

We randomly sample one candidate, and generate a descriptive prompt using the 8-bit quantised (Jacob et al., 2017) LLaVA model (Liu et al., 2023). We perform inpainting using PowerPaint (Zhuang et al., 2024), guiding the inpainting process with the LLaVA-generated descriptive prompt as a positive prompt and a fixed negative prompt to suppress undesired content. We set the fitting degree (which controls how closely the output adheres to the shape of the mask) to 0.6 (0-1), use 45 (0-50) diffusion steps to balance quality and compute efficiency, and apply a guidance scale of 7.5 (0-30) that controls for alignment with the prompt.

We verify resulting images using LLaVA through a binary visual query assessing the presence of the generated candidate object in the image; if rejected, the guidance scale is increased by 7.5, and the process is repeated up to four times until the maximum guidance scale level of 30. If all attempts fail, a new candidate is sampled. This process yields 8 generated images per input (3 each for the low- and medium relatedness conditions, 1 for high relatedness, and 1 for the same-target condition). A second round of filtering is applied using full-precision LLaVA to ensure final quality.

A.2 Spatial encoding & model prompts

KOSMOS-2

KOSMOS-2 encodes spatial regions by discretizing bounding boxes into location tokens and embedding them alongside natural language using `<grounding>` tags.
Prompt:

> <grounding>What is the object in <phrase>this part</phrase> of the image? Answer with the object's name only. No extra text. Answer:

KOSMOS-2 KOSMOS-2 employs a discretized token-based representation of spatial information. The key principle is to transform continuous image coordinates into discrete location tokens via a grid-based partitioning of the image space. Each image is divided into a $P \times P$ grid, and every bounding box is encoded by mapping its top-left and bottom-right corners to location tokens. These tokens are then embedded in the textual input using a structured format that mimics hyperlink annotations, explicitly marking visual references within language. This design enables KOSMOS-

2 to align and jointly reason over linguistic spans and visual regions in a shared sequence modeling framework.

Molmo Molmo grounds language in visual space using a point-based strategy. Instead of bounding boxes, it represents spatial references as single (x, y) coordinates, normalized to a fixed-scale [0,100] grid. Such grounding enables the model to identify, count, or compare objects based on minimal visual supervision. The model interprets these point references directly from the linguistic input, treating them as spatial queries in the generation process.

Molmo

Molmo grounds text in visual space using single-point references normalized to a [0,100] scale. These coordinates are embedded directly in the natural language prompt.
Prompt:

> What is the object at point x = 63, y = 44 of the image? Answer with the object's name only. No extra text.

xGen-MM (BLIP-3) xGen-MM encodes spatial grounding using natural language expressions that describe object locations via bounding box coordinates.

xGen-MM

xGen-MM embeds bounding box coordinates within the prompt using `<bbox>` tags and learns to interpret these as visual region references in natural language form.
Prompt:

> <|user|> <image>What is the object in this part <bbox>[x1, y1][x2, y2]</bbox> of the image? Answer with the object's name only. No extra text.<|end|> <|assistant|>

Qwen2.5-VL Qwen2.5-VL builds upon the Qwen-VL architecture by integrating visual tokens into a large language model using enhanced spatial encoding strategies. It employs bounding box references enclosed in square bracket format $[x_1, y_1, x_2, y_2]$, where coordinates are normalized within the range [0,1]. The spatial references are directly inserted into the natural language prompt, making them interpretable as explicit multimodal instructions.

Qwen2.5-VL

Qwen2.5-VL encodes normalized bounding box coordinates directly in the prompt as $[x_1, y_1, x_2, y_2]$, supporting explicit multimodal alignment through instruction-tuned sequences.
Prompt:

> What is the object in this part of the image [x1, y1, x2, y2]? Answer with the object's name only. No extra text.

LLaVA-OneVision LLaVA-OneVision employs a normalized bounding box encoding strategy embedded directly in the textual prompt in the form `[x1, y1, x2, y2]`, where all values fall within the [0,1] range.

LLaVA-OneVision

LLaVA-OneVision embeds bounding boxes in [0,1] range directly in the prompt, leveraging instruction-tuned pretraining to interpret spatial references as tasks.
Prompt:

> What is the object in this part of the image [x1, y1, x2, y2]? Answer with the object's name only. No extra text.

A.3 Accuracy comparison between dataset labels and modal object names in VISIONS

We use the modal names provided in the VISIONS dataset to compare our cosine-similarity-based accuracy metric under two different ground truths: (1) the dataset labels and (2) the human-generated modal names. This comparison allows us to assess how sensitive the metric is to variation in naming. The contingency tables 4 show a strong alignment between the two ground-truth conditions. Predictions judged incorrect using the dataset label are almost always incorrect using the modal name as well (98.1%), and pre-

Table 4: Contingency table showing the counts and row percentages of agreement and disagreement between accuracy computed with dataset labels (rows) and with modal names (columns), along with the overall correlation and number of divergent cases.

Counts

Accuracy		0	1
Accuracy Modal Names	0	4817	73
	1	95	955

Row Percentages

Accuracy		0	1
Accuracy Modal Names	0	98.1	7.1
	1	1.9	92.9

Pearson correlation: 0.902
Divergent cases: 168 / 5940 (2.83%)

dictions judged correct with the dataset label are likewise correct with the modal name in most cases (92.9%). The low divergence rate (2.83%) and high correlation (0.902) indicate that disagreements between the two ground-truth settings are rare and that the metric captures nearly identical performance patterns across them. Overall, these results suggest that the accuracy metric is robust to synonym variation and is not substantially affected by differences in naming between dataset labels and human-provided modal names.

A.4 Attention over image tokens in LLaVA-OneVision

LLaVA-OneVision handles an image–question pair in four stages:

1. Following the AnyRes visual tokenization strategy, the input image is resized and tiled into $a \times b$ crops of the same resolution. Each view – including the full image – is encoded by the SigLIP SO400M ViT (Zhai et al., 2023) into T tokens, yielding $L = (a \times b + 1)\,T$ tokens in total. If L surpasses the threshold τ, bilinear interpolation reduces the per-crop token count.
2. All visual tokens are projected via a two-layer MLP into the LLM's embedding space;
3. The resulting image tokens are prepended to the tokenized question together with an answer-start marker;
4. The LLM then decodes the answer autoregressively.

In Section 6.2, we restrict our analysis to the visual tokens obtained from the fixed encoding of the full image, omitting any multi-crop tokens.

Below, we detail the process of aggregating attention values and computing the layerwise attention allocation ratio $r_l^{(t)}$ described in Section 6.2.

A.5 Aggregating LLM Attention

We first extract the attention deployed by the LLM decoder over image tokens. Let the attention tensor be $A \in \mathbb{R}^{L\times H\times N\times N}$, where L is the number of layers, H is the number of attention heads, and N is the number of tokens. For each layer l, we compute the mean attention over heads: $\bar{A}_l = \frac{1}{H}\sum_{h=1}^{H} A_{l,h}$. To improve interpretability, we follow a common practice of zeroing out attention directed toward the beginning-of-sequence (BOS) token, which often absorbs a disproportionate amount of attention due to its special role in autoregressive models (Vig and Belinkov, 2019). Specifically, for each attention matrix $\bar{A}_l$, we set the column corresponding to the BOS token to zero, effectively removing all incoming attention to this token. We then normalize each row so that its values sum to one, resulting in a normalized attention matrix $\hat{A}_l \in \mathbb{R}^{N\times N}$.

We aggregate attention across layers to obtain the final LLM attention matrix: $A_{\text{LLM}} = \frac{1}{L}\sum_{l=1}^{L} \hat{A}_l$. Our method distinguishes between prompt tokens and generated tokens in accordance with the autoregressive nature of the model. Prompt tokens receive attention from all subsequent tokens, while generated tokens attend only to preceding tokens. To construct the attention representation, we first process the prompt tokens by computing their attention matrix as described above.

For generated tokens, we apply a similar procedure with two key modifications: (1) only the last row of each attention matrix is retained, corresponding to the attention from the newly generated token, and (2) attention to the BOS token is nullified as before. Each resulting attention vector is

then padded to a common length and stacked with the aggregated prompt token attention matrix.

A.6 Aggregating Attention in Vision Transformers (ViTs)

During our initial experiments, we noticed a tendency where the model applied attention over a specific set of tokens for many images, indicating a model-intrinsic bias. To mitigate such systematic attention biases, we compute a layer-wise average attention map across the dataset and subtract it from each instance's attention map. This adjustment emphasizes instance-specific patterns by removing model-intrinsic tendencies. To avoid resolution-based confounds, we include only images of size 640×640 (which comprise 86% of the dataset).

Per-Instance Layer Aggregation: For each image instance $t \in T$ and layer $l \in L$, we average the attention across heads: $\bar{A}_l^{(t)} = \frac{1}{H} \sum_{h=1}^{H} A_{l,h}^{(t)}$. We then normalize each row of $\bar{A}_l^{(t)}$ so that attention values sum to one, yielding a normalized attention map $A_{\text{norm},l}^{(t)}$.

Dataset-Wide Averaging: To capture general attention patterns, we compute the average normalized attention map across all instances: $A_{\text{avg},l} = \frac{1}{T} \sum_{t=1}^{T} A_{\text{norm},l}^{(t)}$. These maps serve as baselines for interpreting the attention behavior of individual instances. The final per-instance attention map is computed by subtracting the baseline and applying ReLU: $A_{\text{ViT},l}^{(t)} = \text{ReLU}\left(A_{\text{norm},l}^{(t)} - A_{\text{avg},l}\right)$. This operation retains only positive deviations from the dataset-wide mean, highlighting unique attention patterns for each instance. Each column of $A_{\text{ViT},l}^{(t)}$, corresponding to a vision token $j \in J$, is reshaped into a spatial attention map $V_j \in \mathbb{R}^{g \times g}$, where g is the number of image patches per side (e.g., $g = 27$ for SO400M). These maps represent spatial distributions of attention at the selected layer.

A.7 Attention Over Image Regions

To analyze how the model distributes its focus across visual inputs during generation, we compute attention over image regions.

Let $O = \{o_1, o_2, \ldots, o_{|O|-1}\}$ denote the set of the LLM decoder output tokens, excluding the final <|im_end|> token, which is not semantically informative. For each output token $o_i \in O$, we extract the corresponding attention weights $A_{\text{LLM},i,j}$ from the LLM attention matrix, where j indexes the vision tokens. These attention weights are normalised to yield $\hat{A}_{i,j}$.

Each vision token j is associated with a spatial map $V_j \in \mathbb{R}^{g \times g}$, derived from the adjusted ViT attention described earlier. The attention map over the image for output token o_i is then computed as a weighted sum: $A_i^{\text{img}} = \sum_{j \in J} \hat{A}_{i,j} \cdot V_j$. To obtain a global attention map for the image, we average over all output tokens: $A_{\text{final}}^{\text{img}} = \frac{1}{|O|-1} \sum_{i \in O} A_i^{\text{img}}$. This attention map is then upsampled to the original image resolution using nearest-neighbor interpolation for visualization.

A.8 Attention Over Target and Context

To analyze how attention is distributed between the target object and the surrounding context, we extract the bounding box $B = (x_{\min}, y_{\min}, w, h)$ of the target object region. The total attention over the target area is computed as: $a_{\text{target}} = \sum_{(x,y) \in B} A_{\text{final}}^{\text{img}}(x, y)$. Similarly, the attention over the context region, excluding both the target and irrelevant padding areas, is given by: $a_{\text{context}} = \sum_{(x,y) \in C} A_{\text{final}}^{\text{img}}(x, y)$, where C represents the pixels outside the target bounding box but within the valid image region.

The entire process is carried out for each layer of the vision backbone model, for each of which we have a specific aggregated attention representation $A_{\text{ViT},l}$, ensuring that we obtain an attention distribution over the image, and thus over both the target and context, specific to each layer.

A.9 Attention Allocation Ratio

To analyze attention allocation across different levels l of the vision encoder and experimental instances t, we compute the ratio r between the total normalized attention values assigned to the target (a_{target}) and the context (a_{context}): $r_l^{(t)} = \frac{a_{\text{target},l}^{(t)}}{a_{\text{context},l}^{(t)}}$. A value of r closer to 1 indicates a higher allocation of attention to the target region; a value closer to 0 suggests greater attention to the context. Since the context area is larger than the target, we do not expect r to exceed 0.5. Instead, our focus is on analyzing its variations across different experimental conditions.

We then compute the mean ratio value for each encoder layer by averaging across all instances: $\bar{r}_l = \frac{1}{T} \sum_{t=1}^{T} r_l^{(t)}$ where T denotes the total number of instances. This allows us to assess how at-

tention is distributed across different hierarchical levels of the encoder under varying experimental conditions.

B Complementary results

	Congruent images			Incongruent images		
	Coef.	Std.Err.	z	Coef.	Std.Err.	z
Int.	1.744	0.054	32.406*	1.745	0.055	31.902*
Context	-0.087	0.013	-6.56*	-0.178	0.013	-13.518*
Target	-0.105	0.013	-7.866*	-0.094	0.013	-7.152*
N. level	-0.040	0.009	-4.261*	-0.028	0.009	-2.953*
N. level × Con	0.009	0.013	0.667	-0.001	0.013	0.044
N. level × Tar	0.041	0.013	3.030*	0.023	0.013	1.78
Group Var	0.020	0.044		0.020	0.046	

Table 5: Linear mixed effects model on VISIONS: Noise area (Context/Target) and noise level (N. level) on scene-target semantic similarity, for congruent and incongruent conditions. * denotes a significant effect at $p \leq .01$. Conditional R^2 values: 0.423 (congruent) and 0.452 (incongruent).

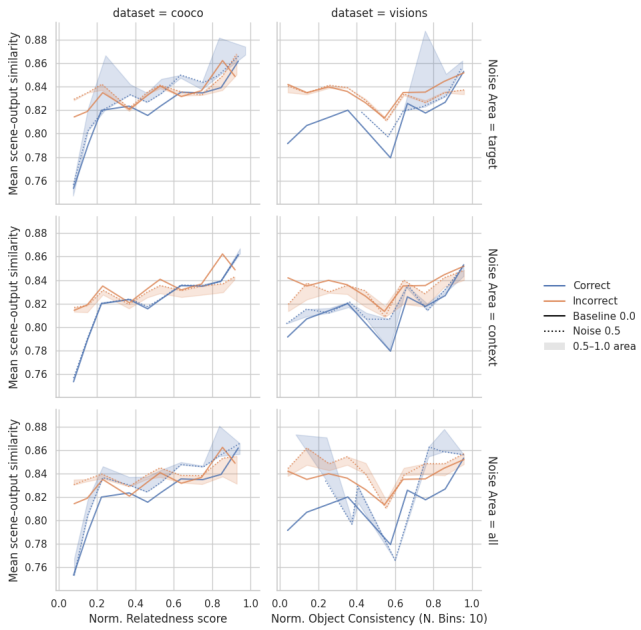
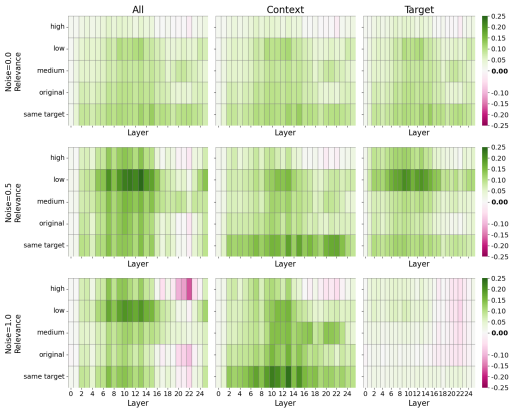
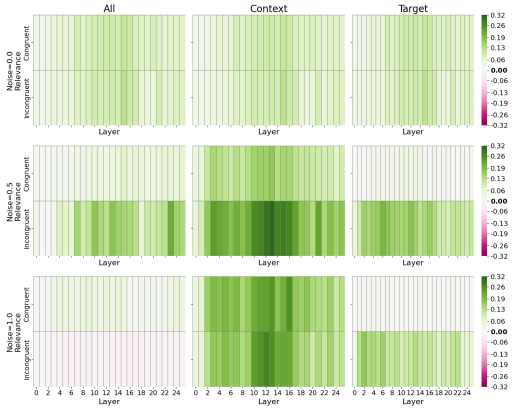


Figure 10: Mean semantic similarity between model outputs and scene labels as a function of binned continuous object–scene relatedness (COOCo) and object consistency (VISIONS), using 10 bins. Columns correspond to datasets; rows to noise–area conditions (target, context, all). Continuous lines represent the baseline (0.0 noise) for correct (blue) and incorrect responses (orange); dotted lines indicate the 0.5-noise condition; shaded regions show the variability introduced when increasing noise from 0.5 to 1.0.



(a) COOCo, $\Delta\bar{r}_l$ (range: ± 0.25)



(b) VISIONS, $\Delta\bar{r}_l$ (range: ± 0.32).

Figure 11: Attention-to-target delta $\Delta\bar{r}_l$, between correct and incorrect predictions across layers, semantic relatedness levels, noise levels, and noise areas. Heatmaps are predominantly green, indicating that when the model correctly recognizes the object, it tends to allocate more attention to the target token compared to when it misclassifies the object.

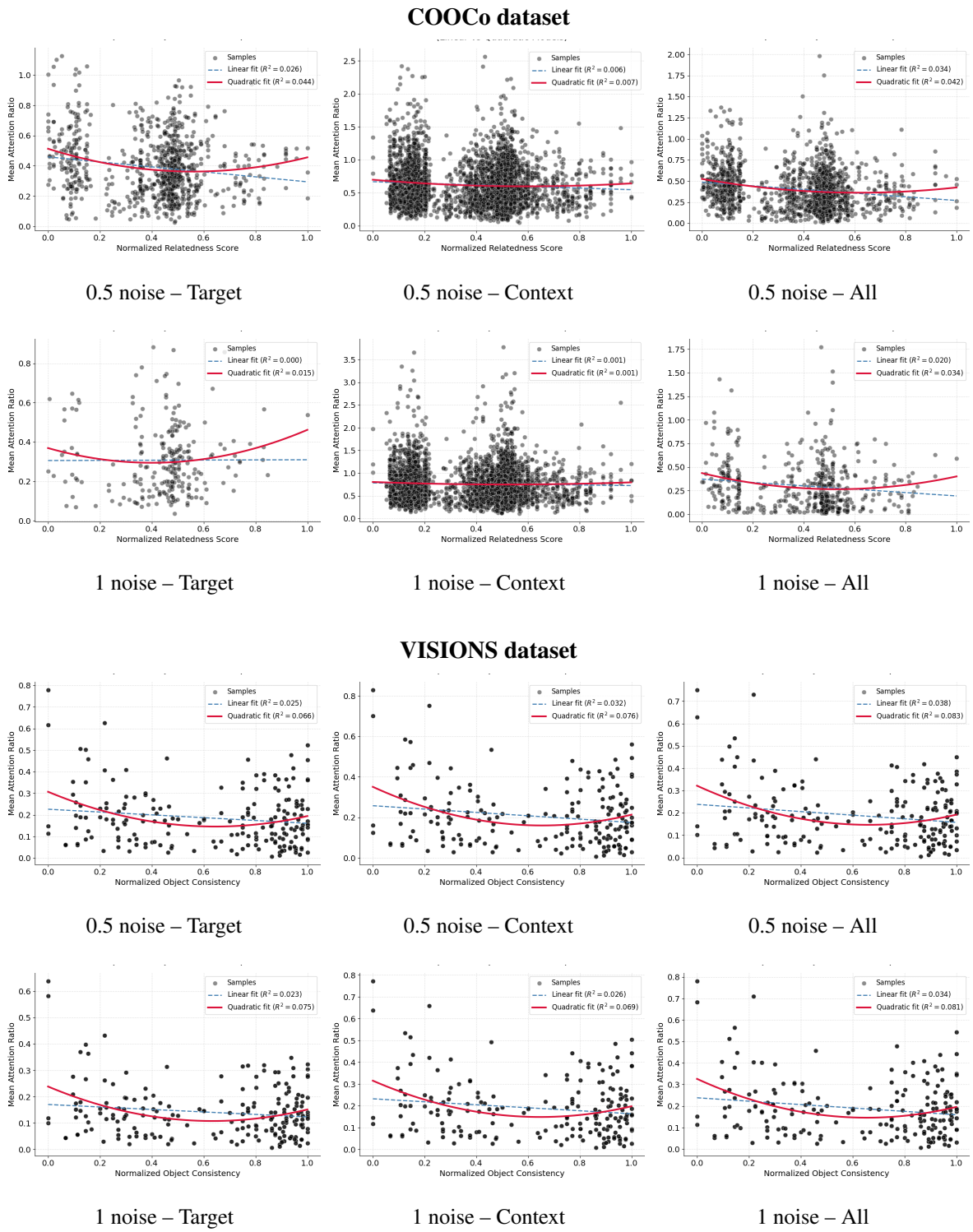


Figure 12: Mean attention-to-target ratio as a function of semantic fit for COOCo and VISIONS, at two noise levels (0.5, 1.0) and noise area conditions (target, context, all).

Table 6: Accuracy results for LLaVA-OneVision-0.5B on the COOCo dataset across relatedness levels, noise intensities, and noise areas.

Rel. Level	Noise Level	Noise Area	Total Samples	Correct Samples	Accuracy (%)
low	0.0	–	111384	41860	37.58
	0.5	all	111384	10816	9.71
		context	111384	41548	37.30
		target	111384	4732	4.25
	1.0	all	111384	3302	2.96
		context	111384	41990	37.70
		target	111384	728	0.65
medium	0.0	–	117494	44954	38.26
	0.5	all	117494	17810	15.16
		context	117494	45448	38.68
		target	117494	11674	9.94
	1.0	all	117494	6084	5.18
		context	117494	43134	36.71
		target	117494	3510	2.99
high	0.0	–	34086	17654	51.79
	0.5	all	34086	8242	24.18
		context	34086	17758	52.10
		target	34086	5252	15.41
	1.0	all	34086	2938	8.62
		context	34086	16692	48.97
		target	34086	1820	5.34
same target	0.0	–	40326	34476	85.49
	0.5	all	40326	23452	58.16
		context	40326	34736	86.14
		target	40326	21008	52.10
	1.0	all	40326	10192	25.27
		context	40326	33722	83.62
		target	40326	8970	22.24
original	0.0	–	41626	34840	83.70
	0.5	all	41626	19734	47.41
		context	41626	37024	88.94
		target	41626	18434	44.28
	1.0	all	41626	8320	19.99
		context	41626	35256	84.70
		target	41626	9490	22.80

Table 7: Accuracy results for LLaVA-OneVision-0.5B on the VISIONS dataset across congruency conditions, noise intensities, and noise areas for incongruent and congruent conditions.

Rel. Level	Noise Level	Noise Area	Total Samples	Correct Samples	Accuracy (%)
Incongruent	0.0	–	8580	1950	22.73
	0.5	All	8580	104	1.21
		Context	8580	1742	20.30
		Target	8580	52	0.61
	1.0	All	8580	104	1.21
		Context	8580	1898	22.12
		Target	8580	52	0.61
Congruent	0.0	–	8580	3250	37.88
	0.5	All	8580	390	4.55
		Context	8580	2626	30.61
		Target	8580	624	7.27
	1.0	All	8580	78	0.91
		Context	8580	2834	33.03
		Target	8580	624	7.27