UNIVERSITÀ DI TRENTO

CiMeC Center for Mind/Brain Sciences

University of Trento

CIMeC, Center for Mind/Brain Sciences

Master's Degree in Cognitive Science - Computational and Theoretical Modelling of Language and Cognition Track

~ · ~

Academic Year 2023–2024

Cognitive and Human-Inspired Evaluation of Vision-Language Models in Scene Understanding

Supervisor and Co-supervisor
Prof. Raffaella Bernardi – University of Trento
Prof. Albert Gatt – Utrecht University

Graduate Student
Filippo Merlo
239305

A chi mi ha sempre spinto a ben fare.

Abstract

Contextual cues are pivotal in human perception and attention, enabling efficient object recognition and scene understanding. Recent advances in vision-language models suggest that context can also support referring expression generation (REG), especially under noisy or partially occluded conditions. Drawing on Melissa Võ's concept of scene grammar, this thesis introduces the *Common Objects Out-of-Context (COOCO)* dataset—a large, carefully curated collection of real and partially generated images in which objects align with or violate scene semantics to varying degrees. We use this resource to evaluate state-of-the-art multimodal models' capacity to harness contextual information for accurate reference expressions, both in typical viewing and under visual degradation. Finally, by analyzing the attention patterns of a top-performing model, we offer insights into how scene information is processed in diverse task settings. Through this work, we aim to deepen our understanding of vision-language models' robustness and flexibility in scene comprehension while bridging a gap between cognitive research on object-context interactions and computational models of visual attention.

Contents

Introduction

Vision-Language Models (VLMs) have demonstrated remarkable progress in integrating visual and linguistic modalities to enhance scene understanding. However, an important yet understudied aspect is their ability to flexibly leverage contextual information during object recognition in Referring Expression Generation (REG) tasks—tasks focused on creating descriptions of entities that enable a listener or reader to uniquely identify them in a given context [25]—particularly in scenarios where object-scene relationships exhibit varying degrees of semantic relatedness. Two key studies have explored this issue: one examining attention allocation in REG models across the different input components (*target*,*location of the target* and *context*)[51], and another investigating the role of scene context as a resilience factor in object description [24]. [51] found that REG models primarily focus on the visual target, paying less attention to contextual features. The study suggests that current models do not effectively integrate contextual cues in a human-like manner, highlighting the need for improved strategies in referential expression generation. Junker et al. [24] demonstrated that contextual information enhances REG model robustness under occlusion and noise, reinforcing its role as a crucial support mechanism. However, they also found that models tend to over-rely on statistical co-occurrence patterns rather than achieving genuine scene understanding, emphasizing the need for better evaluation frameworks that better capture the nuanced interplay between object recognition and scene comprehension. From a cognitive science perspective, research has shown that human perception and attention allocation are strongly influenced by scene context [2, 40, 61, 43]. Humans use contextual priors to efficiently locate and recognize objects in complex environments, often relying on hierarchical scene structures [61]. When objects appear outside their usual context or in unexpected locations, increased attentional resources are deployed to resolve the incongruity [39]. Unlike current VLMs, which often default to statistical correlations, human perception integrates top-down expectations with bottom-up visual features to dynamically adapt to varying scene constraints. This gap highlights the need to investigate whether computational models can approximate human-like flexibility in leveraging context for object understanding. Building upon these insights, this thesis aims to assess how VLMs leverage scene context under varying levels of relatedness and visual degradation. We introduce the *Common Objects Out-of-Context (COOCO)* dataset to systematically manipulate object-scene congruence and evaluate the robustness of multimodal models in diverse scene conditions.

0.1 Research Questions

The goal of this thesis is to investigate how vision-language models process contextual information in scene understanding and referring expression generation. Specifically, we address

the following research questions:

- To what extent do VLMs leverage contextual information in object identification and REG tasks?
- How do these models respond to varying degrees of semantic congruence between objects and their surrounding scenes?
- Does contextual information improve the robustness of REG models under visual noise and occlusion conditions?
- How do state-of-the-art VLMs distribute attention across scene components when processing reference expressions?

0.2 Contributions

This work presents several contributions to the fields of vision-language modeling, cognitive-inspired AI evaluation, and dataset development:

- **Dataset Innovation:** By releasing *COOCO*, we provide the community with a large-scale, precisely annotated resource for analyzing semantic scene violations.
- **Contextual Benchmarking:** We present a cognitively inspired evaluation framework that tests VLMs resilience to occlusions and semantic incongruities.
- **Theoretical Insights:** By linking our evaluations to human studies on scene grammar and attention, we shed light on the nature of contextual representations in deep architectures, bridging cognitive science and AI research.

0.3 Structure of the Thesis

This thesis is structured as follows:

- **Chapter 1: Related Work** - Reviews existing literature on the role of context in human visual perception, cognitive models of attention, and recent advances in multimodal deep learning models.
- **Chapter 2: Proposal** - Outlines our research objectives, hypotheses, and methodology for evaluating VLMs in scene understanding.
- **Chapter 3: Dataset** - Introduces the COOCO dataset, detailing its design, data collection process, and intended applications.
- **Chapter 4: Models** - Describes the VLMs selected for the evaluation.
- **Chapter 5: Experiment** - Explains the experimental design, including the prompt generation strategy, noise conditions, and evaluation metrics.
- **Chapter 6: Results** - Presents empirical findings from our model evaluations, analyzing their performance in different conditions.
- **Chapter 7: Attention Deployment Analysis** - Examines how VLMs distribute attention across scenes.

- **Chapter 8: Analysis** - Discussion of the key findings outlined in the Results and Attention Deployment Analysis Chapters.
- **Conclusion** - Summarizes contributions, limitations, and potential directions for further research.

By systematically evaluating how VLMs utilize contextual information, this study contributes to a deeper understanding of their strengths and limitations in replicating human-like scene perception. The findings have implications for the development of more robust and context-aware vision-language models, fostering advancements in AI systems that interact with complex, real-world environments.

Chapter 1

Related Work

1.1 The Cognitive Perspective

Objects rarely appear in isolation; instead, they exist within environments rich in contextual cues. A chessboard is typically accompanied by chess pieces, a stethoscope is often found alongside medical tools, and a park bench is usually situated near pathways or trees. These associations shape how we perceive and recognize objects in everyday scenes. When encountering a computer monitor on a desk, for instance, we naturally expect to find a keyboard and mouse nearby. Such contextual relationships help guide our expectations, allowing for more efficient visual processing and faster object recognition.

Perceiving and recognizing objects in real-world settings is inherently tied to the surrounding context. Rather than processing objects in isolation, the human visual system leverages scene context to facilitate perception, recognition, and attention allocation [2, 40, 61, 43]. This chapter reviews key findings on the interplay between scenes and objects, highlighting how both global and local scene properties influence object recognition and visual attention. Additionally, we explore the effects of semantic violations, the hierarchical structure of scenes, and the implications for computational models of vision.

1.1.1 The Role of Context in Visual Perception

Our experience with the world provides us with expectations regarding which objects belong in a given scene and where they are likely to be located. In real-world environments, objects do not appear in isolation; they systematically co-occur with other objects and specific surroundings, offering a wealth of contextual associations that the visual system can exploit.

The structure of many real-world scenes is governed by strong configural rules, akin to those that define the organization of individual objects. This concept is illustrated in Figure 1.1. By averaging hundreds of images aligned on frontal faces, a common intensity pattern emerges, revealing a consistent organization of facial features shared across the category of 'face.' Similarly, average images centred on a single object can reveal structured background elements beyond the object itself. For example, a keyboard's average image often includes a monitor and table in the background, despite the images not being explicitly constrained to contain these objects. In contrast, a fire hydrant's background is less distinct, yet its necessity for ground support leads to an emergent ground plane in the average image. The presence of a particular object thus constrains both the identity and spatial arrangement of surrounding objects, a property that is likely leveraged by the visual system.

Figure 1.1: The structured relationship between objects and their backgrounds. Each image represents an average of hundreds of pictures containing a central object (a face, keyboard, or fire hydrant) at a fixed scale and pose. Source: [40]

Cognitive research has shown that object recognition is facilitated when objects appear in familiar, semantically coherent environments. In contrast, objects that are semantically inconsistent with their surroundings tend to attract additional attention due to their unexpected nature [39]. Moreover, when an object's recognition cannot be swiftly achieved based solely on its physical attributes, contextual information often provides more crucial input than the object's intrinsic properties. A similar effect occurs when objects are introduced following a contextual scene: if they visually resemble an object typically associated with that context, they are more likely to be misrecognized [2]. This phenomenon is illustrated in Figure 1.2, where the same object is perceived differently depending on the surrounding context. In the left panel, the object is interpreted as a hairdryer, while in the right panel, it is perceived as a drill. In both cases, contextual information plays a decisive role in resolving ambiguity.

Figure 1.2: The same object is perceived differently based on contextual cues: a hairdryer in the left panel and a drill in the right panel. Source: [2]

1.1.2 Global vs. Local Contextual Effects

While it is evident that scene contexts influence object processing, the specific "ingredients" of a scene that drive this modulation remain unclear. What aspects of a scene are sufficient to produce the reported consistency effect? Scene context could influence object processing in at least two distinct ways: through global effects arising from broad scene properties or more localized effects due to the immediate surroundings of a specific object.

To investigate the influence of global image statistics on object processing, researchers have examined how scene textures, visual representations that retain global statistical properties but lack spatial layout information, modulate object recognition. These textures preserve summary statistics based on low-level visual features but discard spatial configuration details. While semantic relationships between objects and their surroundings are known to facilitate recognition, this study explored whether even low-level image properties contribute to this effect. Using textures derived from real-world scenes [28] or close-ups of materials (e.g., tiles) [27], researchers found that participants categorized these textures more accurately than colour-matched controls, though original scenes still yielded the highest performance. When recognizing briefly presented objects, participants exhibited a strong advantage for contextually consistent objects in original scenes. Although this effect was weaker for scene textures or close-ups of materials, it remained significant. These findings suggest that low-level visual features play a role in object recognition, even in the absence of explicit semantic structure. Figure 1.3 illustrates the stimulus set used in this study, displaying objects superimposed on four different background conditions: original scenes (left column), scene textures that retain summary statistics while discarding global shape information (top middle), close-ups of materials (bottom middle), and colour-matched controls (right).

Beyond global scene properties, localized contextual information plays a crucial role in object recognition. Objects themselves can evoke scene context representations, facilitating the identification of other elements within a scene. Co-occurring objects provide strong associative cues, shaping expectations about what is likely to appear nearby. Through repeated exposure, we develop an understanding of not only the typical placement of target objects but also the expected positioning of distractors, enhancing search efficiency.

Both local and global scene information contribute to object recognition and search efficiency, highlighting the importance of understanding their interplay in human perception.

The following section will explore in greater detail how local context influences object processing, introducing a recently proposed framework that describes scene structure through the concept of visual grammar.

1.1.3 The Visual Grammar of the Scene

A new approach to framing visual scenes has recently been proposed by [61], advocating a shift away from traditional vision research, which has primarily relied on artificial and simplified stimuli to study fundamental perceptual mechanisms. Instead, this perspective emphasizes the study of real-world perception, where structured, meaningful scenes provide crucial context for object recognition and attention.

The concept of scene grammar, the structured relationships between objects within a scene, has been introduced as a means to bridge the gap between controlled experimental settings and real-world vision. Just as words form structured sentences, visual scenes follow rule-governed configurations: objects are not randomly arranged but tend to appear in predictable spatial and functional relationships. These regularities, learned through repeated exposure (e.g., a pot on a stove in a kitchen), allow us to generate expectations about a scene's structure. Consequently, perception is not merely reactive but proactive, as object contexts sharpen our predictions about elements that may not yet be fully visible [2, 61, 43].

When these structural expectations are violated, two primary types of inconsistencies arise. Semantic violations occur when an object is incongruent with the overall meaning of the scene (e.g., a piece of cheese in a bathroom). Syntactic violations, on the other hand, involve objects that are thematically appropriate but positioned in unexpected ways (e.g.,

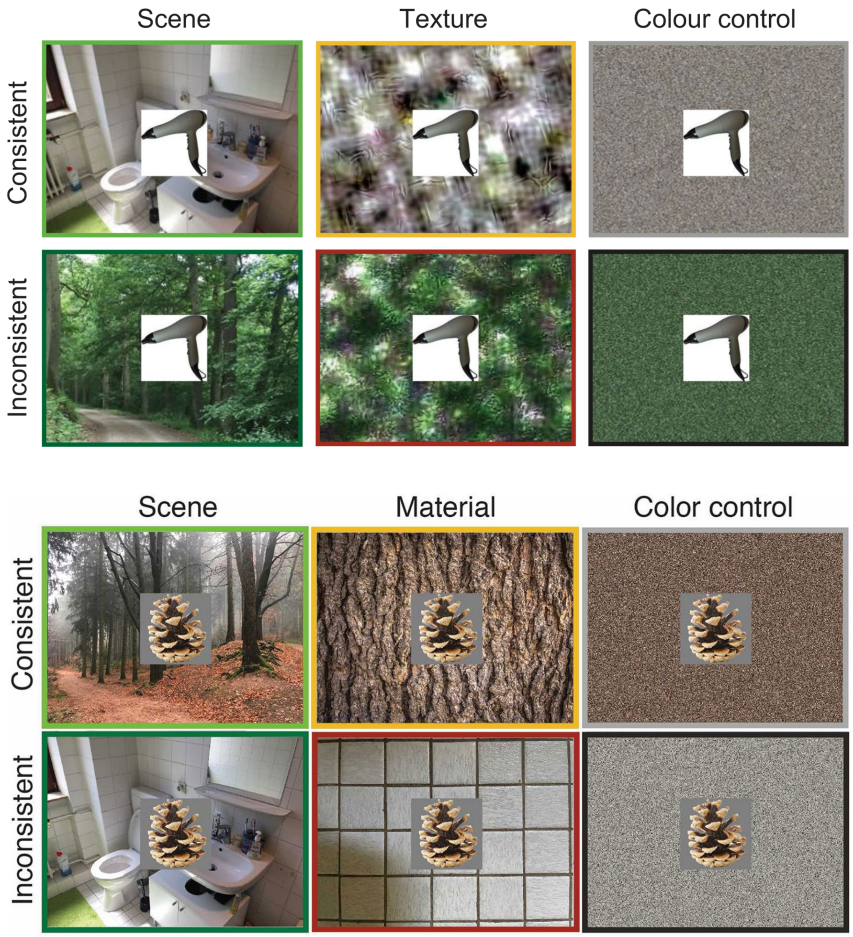


Figure 1.3: Example of a stimulus set with objects superimposed on different background conditions: original scenes (left column), scene textures preserving summary statistics but lacking global shape information (top middle), close-ups of materials (bottom middle), and color controls (right). Source: [28, 27].

a toilet paper roll inside the shower).

Hierarchical Scene Organization

Scenes are structured hierarchically, with certain objects acting as anchors that define the spatial organization of a space. Anchors, such as a stove in a kitchen or a sink in a bathroom, serve as reference points around which related objects are arranged.

Anchor objects differ from merely large or diagnostic objects in that they establish spatial relationships with other objects. While diagnostic objects indicate the type of scene, anchors provide predictions about object locations. For instance, a toilet brush suggests a bathroom but does not predict the exact location of toilet paper. Figure 1.5 illustrates a proposed hierarchical organization of a bathroom scene, where key anchor objects (e.g., a shower, a toilet, and a sink) predict the locations of associated items (e.g., shampoo in the shower, a toothbrush on the sink, and toilet paper next to the toilet).

Figure 1.4: In both language as well as scenes, the 'grammar' of the input allows us to fill in the missing information (ball). Source: [3]

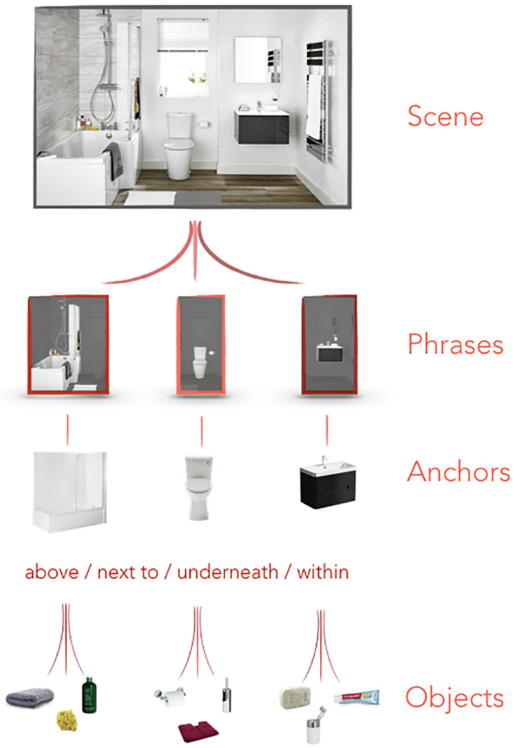


Figure 1.5: Proposed hierarchical organization of a bathroom scene that includes three phrases that again consist of one anchor each (e.g. a shower, a toilet and a sink) that predict the locations of other objects (e.g. the shampoo is in the shower, the toothbrush on top of the sink, the toilet paper next to the toilet, etc.). Source: [61]

In [3], the concept of anchors has been operationalized by quantifying the spatial relationship between objects in the scene database SUN [66] with human-annotated objects. The anchor selection in this study is validated using four determinants:

1. **Object Pair Frequency (OPF)**: Measures how often an anchor co-occurs with a target object within a scene category. It reflects the likelihood of their spatial

association but does not guarantee exclusive presence.

2. **Object Mean Distance (OMD)**: Captures the average spatial proximity between an anchor and its associated objects, normalized for image size. Anchors typically appear closer to their targets than swapped alternatives.

3. **Object Spatial Vertical Variance (OSV)**: Represents the consistency of vertical spatial arrangements between an anchor and other objects across scenes. Lower variance (higher inverse variance) suggests stable positioning, such as a toothbrush consistently appearing above a sink.

4. **Object Cluster Frequency (OCF)**: Derived from clustering algorithms, it indicates how often an anchor appears as the largest object in a cluster. A higher OCF suggests a stronger anchoring role within the scene.

These measures quantify an object's anchoring properties, providing a systematic way to assess its influence on spatial organization.

Understanding how scenes are composed and how objects are organized within them is crucial for predicting object locations. In indoor scenes, objects tend to cluster around anchors, forming meaningful subgroups or phrases (e.g., a sink, shower, and toilet forming a bathroom phrase).

This knowledge enhances search efficiency; for instance, when searching for shampoo, one can focus on the shower rather than other bathroom subunits. Similarly, scene composition speeds up object recognition; if an object is glimpsed atop a stove, prior knowledge of typical stove-associated objects aids identification [2, 61].

Research has shown that anchor objects facilitate object search and recognition by providing spatial cues. In a series of three experiments, [3] provide initial evidence for the role of anchor objects in guiding search through scenes. To investigate this, they generated three-dimensional (3-D) images of scenes in which a critical anchor object (e.g., the shower) was swapped out for a similar, semantically consistent surrogate object (e.g., a cupboard), which did not serve as an anchor for the current target (e.g., a towel). An example of the generated images in the different conditions is depicted in figure 1.6. In the experiments, participants searched freely for these targets with or without a preview of the target-absent scene before receiving the target probe and beginning their search with a gaze-contingent window. They found a consistent effect of anchors on eye movements. When participants' search was restricted to a gaze-contingent window, response times were significantly affected. Participants searching for local objects exhibited prolonged search times and increased fixation dispersion when anchors were absent. These findings suggest that anchors play a crucial role in guiding efficient search, scene perception, and object identification.

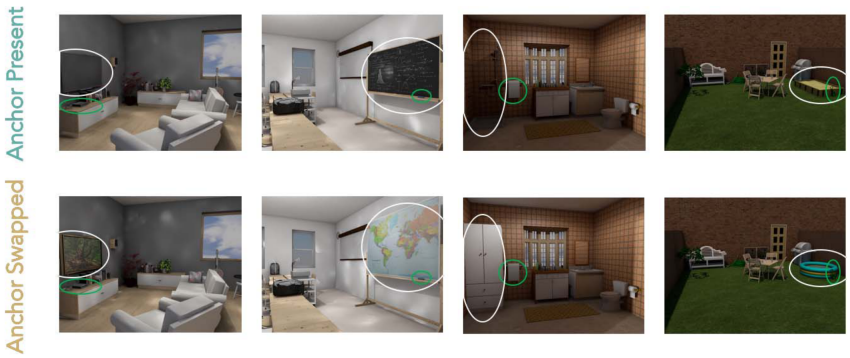


Figure 1.6: Examples of the 3-D rendered scenes used in the experiments. Targets are circled in green, and anchors—or their swapped counterparts—are circled in white. The top row shows the anchor-present trials (from left to right: television, blackboard, shower, sand box) and the bottom row shows the swapped images (from left to right: picture, map, cupboard, swimming pool). Source: [3]

A better understanding of the role of anchor objects in object search could enhance models' ability to predict eye movements and improve the identification of small or occluded objects. Existing models, like the contextual guidance model [57], predict broad, unconstrained search areas. However, a model incorporating spatial predictions based on anchor objects could significantly narrow predicted search areas by first identifying anchors and then restricting searches to locations relative to them. This multi-stage strategy, which humans naturally use, could be beneficial for machine vision systems [62, 61].

The Meaning of the Scene

Research has consistently shown that objects appearing in semantically consistent scenes are recognized more quickly and accurately than those in incongruent contexts [28, 27]. When an object deviates from expected scene semantics (e.g., a fire extinguisher in a kitchen), it tends to attract more visual attention, leading to prolonged fixations and delayed recognition [39]. Notably, semantic consistency appears to function along a continuum rather than as a binary distinction, with both highly consistent and highly inconsistent objects receiving more attention than moderately consistent ones [9].

However, semantic incongruency can sometimes enhance task performance. For instance, [53] investigated the impact of semantic congruency on object perception in visual scenes through two behavioural experiments. In the first, participants performed a change-detection task in which they viewed alternating images of a scene with and without a target object and had to identify the change. Results indicated that detecting changes was more difficult when the changing object was congruent with the scene, suggesting that congruent objects were less perceptually salient. The second experiment examined whether this effect extended to object identification: after briefly viewing a scene, participants had to choose which of two similar objects had appeared in it. Accuracy was lower, and response times were slower when the target object was congruent with the scene, indicating that congruency impaired recognition. These findings suggest that objects fitting well within a scene are processed less distinctly than incongruent ones, impacting both visual change detection and object identification.

To further explore how human attention is influenced by scene semantics, some studies

have employed semantic maps [19, 9], a tool we will review in the next section. Building on this approach, we developed a new dataset specifically designed to examine semantic violations.

1.1.4 Visual Attention, Semantic Maps and Semantic Violations

One of the key questions in the field of scene perception has been what determines where and when we attend during scene viewing.

Traditional models of attentional guidance emphasize image salience, where attention is drawn to visually distinctive regions based on low-level features like luminance, colour, and edges. These models rely on saliency maps to predict eye movement patterns [19].

Recently, Henderson and colleagues have started directly comparing the influences of meaning and image salience on attentional guidance in real-world scenes. To compare these influences quantitatively, they developed meaning maps, which represent the spatial distribution of semantic informativeness. These maps are created through crowd-sourced ratings of scene patches, highlighting meaningful regions independent of visual features. Meaning maps allow direct comparisons between semantic informativeness and image salience in guiding attention [19].

More recent research has introduced concept maps as a novel approach to studying attentional guidance in real-world scenes. Unlike traditional meaning maps derived from crowd-sourced ratings, concept maps are generated using vector-space semantic models. Hayes and Henderson [17] constructed these maps by computing the semantic similarity between objects in a scene using ConceptNet Numberbatch [54], which integrates distributional (Word2Vec, GloVe) and knowledge-based (ConceptNet) representations. Each object's similarity to all other objects and the overall scene category was averaged, then mapped to its spatial location, and finally smoothed with a Gaussian filter to create a semantic density representation. Analysis of eye movements revealed a strong link between concept map values and visual attention. An example of the images and concept maps used in this study is presented in Figure 1.7. Regions with higher semantic similarity were more likely to be fixated. Importantly, this effect could not be fully explained by low-level visual salience, reinforcing the idea that semantic knowledge plays a fundamental role in attentional guidance in real-world scenes.

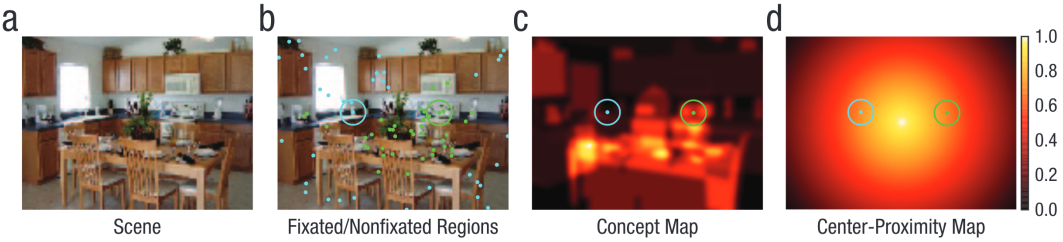


Figure 1.7: Example scene (a) with fixated and non-fixated regions for one participant (b), along with corresponding concept map values (c) and center-proximity map values (d). In (b), green dots mark fixation locations, while cyan dots indicate randomly sampled non-fixated regions. These locations were used to compute mean ConceptNet similarity (c) and center-proximity (d) values. The heat maps represent cosine similarity (c) and scaled centre proximity (d). Source: [17]

In line with these findings, Damiano et al. [9] investigated the semantic-inconsistency

effect in visual scene perception employing the same continuous measure of inconsistency based on linguistic-semantic similarity, the same employed in the concept mapping approach. Their study employed eye-tracking to examine how participants visually explored real-world indoor scenes from the SCEGRAM database [39] (detailed in Section 3.1.3), each containing an object varying in its degree of semantic fit with the scene category. Semantic consistency was quantified using Latent Semantic Analysis (LSA) scores derived from ConceptNet Numberbatch embeddings, allowing for a graded assessment of an object's relation to the overall scene. Specifically, two measures were computed: (1) object–category similarity, capturing how closely an object aligns with the broader scene semantics, and (2) object–object similarity, assessing its semantic coherence with other objects in the scene. The results revealed a negative correlation between LSA scores and the semantic-inconsistency effect, with objects of lower semantic similarity (higher inconsistency) attracting more visual attention. Notably, the study uncovered a U-shaped relationship, where both highly consistent and highly inconsistent objects received more fixations and longer gaze durations than those with intermediate consistency. This pattern aligns with prior research on meaning and concept maps, reinforcing the notion that visual attention is modulated not only by low-level salience but also by semantic expectations. By modeling semantic guidance as a continuous variable using linguistic representations, this work provides further evidence that attention in real-world scene perception is shaped by the graded interplay of meaning and contextual coherence.

1.2 Context Aware Models

Several computational models have attempted to model how context guides attention in object search and enhances object recognition. These models integrate both local and global visual features to refine search processes, aiming to mimic human-like efficiency in identifying relevant objects within a scene.

1.2.1 The Contextual Guidance Model

The Contextual Guidance Model [57] posits that global scene properties constrain attention, while local object relationships refine search mechanisms. This model introduces an innovative approach to attentional guidance by utilizing a global scene context. It consists of two parallel pathways: one computing local features (saliency) and the other computing global (scene-centered) features. By integrating bottom-up saliency, scene context, and top-down mechanisms, the model predicts regions likely to be fixated by human observers performing natural search tasks in real-world scenes.

The image processing in this model occurs in two parallel pathways:

- **Local Pathway**: Represents each spatial location independently, computing image saliency and performing object recognition based on local appearance.
- **Global Pathway**: Extracts holistic global statistics from the image for scene recognition, providing contextual information about expected object locations.

Both pathways share an initial stage in which the image is filtered by multiscale-oriented filters. The global pathway then informs the local pathway about likely object locations, improving search efficiency by narrowing down the search space before detailed analysis.

Figure 1.8 illustrates the architecture of the Contextual Guidance Model, demonstrating the integration of local and global pathways for guiding visual attention.

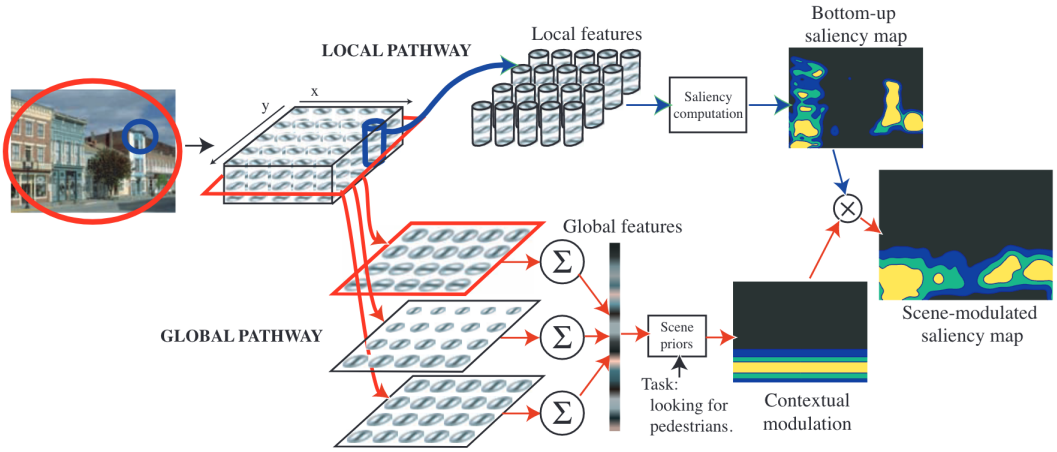


Figure 1.8: An illustration of the Contextual Guidance Model. Source: [57]

1.2.2 Context-aware Recognition Transformer Network (CRTNet)

Bomatter et al. [4] introduce the Out-of-Context Dataset (OCD), a synthetic dataset generated using a Unity-based 3D simulation engine to systematically manipulate scene context, including gravity, object co-occurrences, and relative sizes. The authors conducted psychophysics experiments to establish human recognition benchmarks under these conditions, comparing results with state-of-the-art computer vision models. Their findings indicate that existing recognition models struggle with atypical contextual cues, leading to significant performance degradation.

To address these challenges, the authors propose the Context-aware Recognition Transformer Network (CRTNet), a model that integrates object and contextual information via a dual-stream transformer architecture. CRTNet processes both object and context streams independently before fusing them using multi-head attention layers. The model includes a confidence-weighting mechanism to modulate reliance on contextual cues, ensuring robustness in recognizing objects placed in out-of-context settings.

CRTNet consists of three main modules:

- **Feature Extraction**: Two separate convolutional neural networks (CNNs) extract features from the cropped target object and its surrounding context. These CNNs are based on DenseNet and are pre-trained on ImageNet before fine-tuning.
- **Integration of Context and Target Information**: The extracted feature maps are tokenized, with context and object embeddings fed into a transformer decoder stack. Multi-head attention layers enable hierarchical reasoning, progressively integrating contextual information with object features.
- **Confidence-Modulated Classification**: CRTNet generates two independent predictions: one based solely on the target object and another incorporating contextual cues. A confidence-weighting mechanism determines the final classification decision, prioritizing object-based predictions in ambiguous contexts.

Figure 1.9 presents an overview of the CRTNet architecture, illustrating how the model processes both object and context features to enhance classification performance.

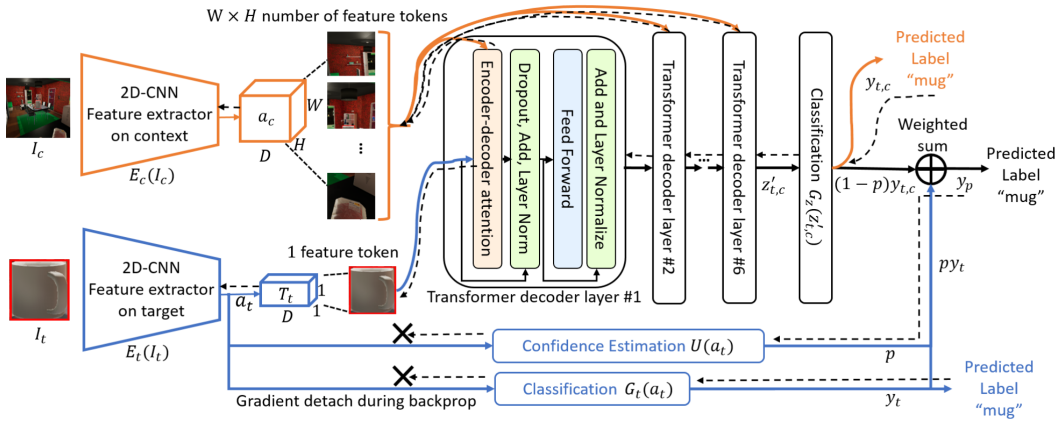


Figure 1.9: Architecture overview of the Context-aware Recognition Transformer Network (CRTNet). CRTNet consists of three main modules: feature extraction, integration of context and target information, and confidence-modulated classification. The model takes the cropped target object I_t and the entire context image I_c as inputs, extracts their respective features, and integrates the information through transformer decoder layers. CRTNet also estimates a confidence score for recognizing the target object based solely on object features, modulating the contributions of y_t and $y_{t,c}$ to the final prediction y_p. The dashed lines in the backward direction denote gradient flows during backpropagation, while the black crosses indicate points where gradient updates stop. Source: [4]

1.2.3 The Target and Context-aware Transformer (TCT)

The Target and Context-aware Transformer (TCT) [13] is a biologically inspired visual search model that integrates both target-driven and contextual modulations to guide self-attention in a Vision Transformer (ViT) [14] framework. TCT modifies a pre-trained ViT by repurposing its self-attention layers into Target and Context-aware Attention Blocks (TCABs). Each TCAB independently modulates self-attention using distinct target and context feature representations:

- **Target Modulation**: Computes patch-wise local relevance between the target and search images. The model extracts target features by mapping the query representation of the target image to search image queries, forming a relevance-driven target modulation matrix. This process applies cross-attention between the target and search scene representations, generating a binary target-search relevance map that modulates self-attention weights.

- **Contextual Modulation**: Derived from a modified Context-aware Recognition Transformer Network (CRTNet) [4], which captures statistical co-occurrences between objects and their surroundings.

TCT operates iteratively, generating sequential fixations based on the maxima of its final attention map:

1. The model selects the highest activation point in the attention map as the next fixation centre.

2. If the fixated area overlaps with the target bounding box, the search terminates successfully.

3. Otherwise, Inhibition-of-Return (IOR) suppresses previously attended locations to prevent redundant fixations.

4. The process continues iteratively until the target is located, closely mimicking human-like visual search behaviours.

By leveraging ViT's multi-head self-attention mechanism alongside human-inspired attention modulations, TCT achieves robust and generalizable performance in naturalistic visual search tasks. Notably, it demonstrates superior search efficiency and zero-shot generalization to novel objects without additional fine-tuning. The modular design of TCT allows for seamless integration with other vision models, making it highly adaptable for applications requiring efficient target localization in varying environmental conditions.

Figure 1.10 illustrates the TCT architecture and its fixation selection mechanism.

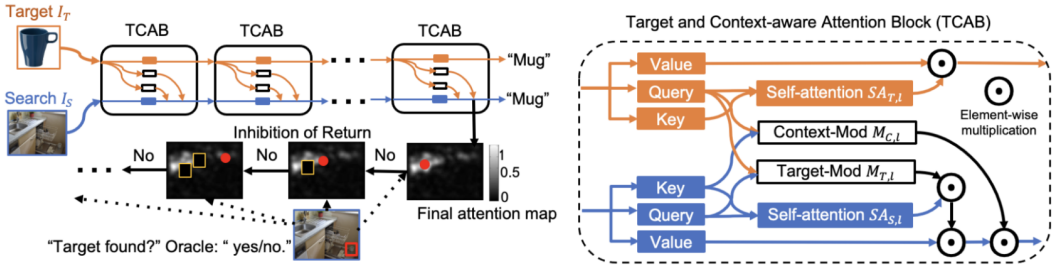


Figure 1.10: Architecture of the Target and Context-aware Transformer (TCT). TCT takes in a target object I_T and a search scene I_S and extracts feature representations of the target and the context independently. The extracted features are applied onto each Target and Context-aware Attention Block (TCAB) in the form of target modulation $M_{T,l}$ and context modulation $M_{C,l}$, guiding attention and producing a final attention map. The model predicts fixations by selecting the maxima of the attention map. If the fixated area overlaps with the target bounding box, the search process ends; otherwise, inhibition-of-return (IOR) suppresses previous fixations. The process repeats until the target is found. Red dots denote predicted eye fixations. Source: [13]

1.3 REG and Context

Referring Expression Generation (REG) is the task of producing a linguistic description that allows a listener to identify a referent within a given context. Historically, REG has been approached through symbolic representations, where objects are characterized by structured attribute sets, and context is defined in terms of high-level abstract properties. More recently, however, research has shifted towards visual REG, where context is represented through raw visual data extracted from images. This transition introduces fundamental differences in how context is conceived, represented, and utilized within REG models [50].

In symbolic REG, context has traditionally been framed as a set of constraints that influence content determination, ensuring that descriptions are both adequate (distinguishing the referent from distractors) and efficient (avoiding redundant properties) [25]. Contextual integration in these approaches can be categorized into different types: distractor context,

which exerts a negative force by ruling out competing entities; perspective cues, which serve a positive function by guiding similarity-based descriptions; and relational context, where landmarks help generate spatially grounded expressions. Additionally, the prominence or salience of context objects modulates the extent to which these types of context exert semantic influence [50].

Visual REG, in contrast, directly processes raw perceptual inputs, mapping them to linguistic descriptions via neural models [69, 35]. This shift alters how context is utilized: rather than relying on explicit symbolic representations, visual REG models extract learned features from convolutional networks or transformer-based attention mechanisms. Context integration in these models primarily follows two paradigms: (1) distractor-focused processing, where visual comparisons between the referent and co-occurring objects enhance discriminability, and (2) scene context as a facilitative resource, where global contextual information aids in recognizing and describing referents, even under perceptual uncertainty. While distractor-focused reasoning remains central to visual REG, additional forms of contextual integration, such as leveraging global scene features for object recognition, remain underexplored [50].

Recognizing objects is a crucial step in visual REG, yet it becomes challenging when visual information is degraded, such as in cases of small object sizes or occlusions. Research in cognitive science and computer vision suggests that contextual information can aid object recognition and categorization across various tasks. Scene context provides valuable cues for disambiguating hard-to-recognize objects. Specifically, scene type can serve as a prior expectation, as certain objects are more commonly found in specific environments. These expectations can be derived from global scene features (e.g., the overall gist of the scene) [28, 27] or local cues (e.g., commonly co-occurring objects) [3]. Additionally, object co-occurrence patterns within a scene can enhance recognition, either by treating scenes as structured collections of objects or by leveraging nearby anchor objects [3]. However, despite its potential, scene context has yet to be systematically integrated into visual REG models [50].

Building on these ideas, the following two studies empirically investigate how REG models differentially attend to targets, locations, and contextual features [51] and how the context may be used as a resilience factor [24]. Together, these studies underscore the need for a more nuanced understanding of how context shapes the referential strategies employed by computational models of REG.

1.3.1 Attention Allocation over Input Partitions

In this study [51] investigate how a Transformer-based Referring Expression Generation (REG) model allocates attention to different input partitions: the visual target (V_t), its location (Loc_t), and the surrounding visual context (V_c). In Figure 1.11 a representation of the three different input partitions is shown. The model processes concatenated feature vectors consisting of these three components, with visual features extracted using ResNet-101. The study is conducted on the RefCOCO and RefCOCO+ [69] datasets, where RefCOCO+ excludes explicit location-based terms like "left" and "right." During training, the model is optimized using Cross Entropy Loss, and its generation quality is evaluated with BLEU [41], CIDEr [59], and METEOR [1] scores.

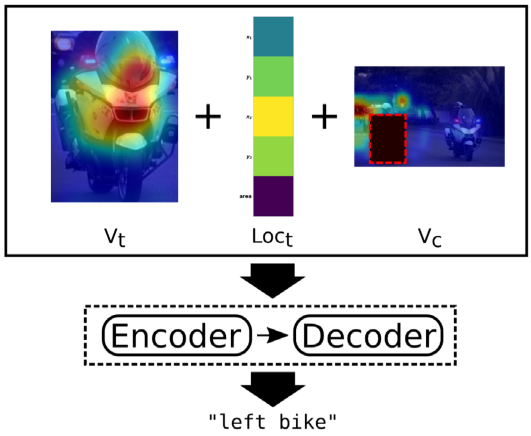


Figure 1.11: Input for our REG model. Input vectors are concatenations of visual (Vt) and location (Loct) features for targets and visual context features (Vc). We examine the relative attention weights of each partition. Source: [51]

To analyze attention distribution, the authors examine self-attention in the encoder and cross-attention in the decoder, computing cumulative attention weights directed to each input partition. Results indicate a strong bias toward the target object, especially in head nouns, whereas subordinate nouns (e.g. "apple" in "apple tree") receive more attention to context. Although location features receive the least raw attention, they become more relevant when normalized for dimensionality, particularly in RefCOCO+. The decoder shows greater variability in attention allocation, suggesting that contextual weighting depends on the generated token type. While the model successfully integrates contextual information in linguistically meaningful ways, it is not explicitly optimized for pragmatic informativeness, limiting its ability to emphasize distinguishing features. They then highlight that work should explore layer-wise attention patterns.

1.3.2 Resilience through Scene Context

In this study [24] investigate the role of scene context in Referring Expression Generation (REG), shifting the focus from a distractor-based context that pressures the speaker to differentiate the expression, to its potential as a resource that enhances model resilience and facilitates object description. The authors hypothesize that contextual information makes REG models more robust, particularly in recognizing object types even under adverse conditions such as occlusion or noise. To test this, they train and evaluate Transformer-based REG models using target representations artificially obscured with noise at varying levels, assessing how different visual contexts impact performance. The results indicate that even simple scene context significantly enhances model resilience, enabling accurate referent type identification even when direct visual information about the target is entirely unavailable.

The experimental setup involves training Transformer-based REG models, including a custom-trained model (TRF) and a fine-tuned pre-trained model (CC). These models incorporate different target and context representations, including visual (bounding box contents and spatial location) and symbolic (scene summaries). Context-enhanced variants process either global image embeddings (TRF_{vis}, CC_{vis}) or symbolic scene summaries (TRF_{sym}, CC_{sym}). To assess robustness, target representations are perturbed with varying noise levels (0.0, 0.5, 1.0), simulating occlusions. TRF is based on ResNet-encoded target and context

features passed to an encoder-decoder transformer, while CC adapts ClipCap [38], mapping visual inputs into GPT-2 [46] prefix embeddings. Both architectures demonstrate that context significantly enhances resilience in REG, particularly in object type identification under occlusion.

noise 0.0	TRF_{tgt}	red van (A)
	TRF_{vis}	red truck (A)
	TRF_{sym}	red truck (A)
noise 1.0	TRF_{tgt}	left elephant (F)
	TRF_{vis}	white truck (A)
	TRF_{sym}	car on left (A)

Figure 1.12: Example from RefCOCO (displayed with noise level 0.5) with generated expressions and human judgments. Visual or symbolic scene context allows to identify even fully occluded targets (noise 1.0). Source: [24]

These models are evaluated using standard automatic quality metrics such as BLEU [41] and CIDEr [59], as well as human judgments assessing the validity of generated expressions.

The results demonstrate that models utilizing contextual information outperform target-only variants, particularly in challenging conditions where visual target information is compromised. Symbolic and visual contexts both contribute significantly to performance. A correlation analysis between identification accuracy and the presence of similar objects in the context confirms that models exploit object co-occurrence patterns to compensate for missing visual information. Attention allocation analyses further show that models shift their focus toward scene context when target information is obscured, indicating an adaptation mechanism that enhances robustness. Qualitative examples highlight cases where models either successfully leverage context for correct referent identification or erroneously copy object types from the scene, suggesting reliance on statistical regularities rather than true scene understanding. The study's findings challenge traditional REG paradigms that emphasize context primarily as a source of distractors, instead revealing its role as a crucial support mechanism in generating accurate and informative descriptions. However, the observed copying strategies suggest that existing datasets like RefCOCO may not fully capture the complexities of real-world visual-linguistic interactions.

The authors advocate for further research integrating insights from perceptual psychology and vision-language generation, aiming to refine REG models toward more human-like scene comprehension and description capabilities. This study contributes to understanding how multimodal models leverage contextual information and underscores the need for more diverse datasets to explore the nuances of scene-based language generation.

Chapter 2

Proposal

Recent advances in vision-language models (VLMs) have led to impressive gains in integrating visual and linguistic information for tasks such as image captioning, visual question answering, and referring expression generation (REG). Nevertheless, a persistent challenge lies in determining how effectively these models leverage contextual cues within a scene. In cognitive science, it is well-established that humans rely on scene context for object recognition and visual attention, particularly under degraded or ambiguous viewing conditions. While current VLMs exhibit strong pattern-matching abilities, they often fail to replicate the adaptive, context-driven strategies characteristic of human perception.

In this proposal, we present a framework for examining how VLMs incorporate contextual information in real-world scenes by drawing on insights from cognitive science and state-of-the-art multimodal architectures. We introduce a novel dataset: *Common Objects Out-of-Context* (COOCO), designed to probe VLMs' capacity to process semantic congruence between objects and their surrounding scenes. By manipulating object-scene relatedness and applying various noise levels to different areas of the visual input, we aim to determine whether context serves as a robust support mechanism or merely as a distractor. Ultimately, we seek to establish a cognitively grounded evaluation benchmark that advances VLMs toward more human-like scene understanding.

2.1 Objectives and Hypotheses

2.1.1 Research Questions

Building on previous findings that underscore the importance of context in human visual perception, this proposal addresses four key questions:

1. **Context Utilization:** To what extent do vision-language models actually employ contextual cues for object identification and description?
2. **Semantic Violations:** How do models respond to semantic incongruities (e.g., objects that exhibit low scene–object relatedness)?
3. **Noise and Robustness:** Does contextual information enhance robustness under visual distortions (e.g., noise, occlusion) in a manner akin to human resilience?
4. **Attention Allocation:** In the presence of semantic violations and degraded targets, how is attention distributed in advanced multimodal models?

2.1.2 Hypotheses

Drawing on cognitive science research into context-driven perception and prior work on referring expression generation, we propose the following hypothesis:

- **Context as Facilitator:** VLMs utilize contextual cues to identify objects and generate referring expressions, particularly when the target region is moderately to severely distorted or noisy. **If this holds true, we expect to see the following effects:**

  1. **Performance Drop in Low-Relatedness Occlusions:** When the target object is occluded and exhibits low semantic congruence with the scene, performance should decline markedly because contextual cues cannot fully compensate for missing visual information. By contrast, this decline is expected to be less severe when objects are highly or moderately related to the scene.
  2. **Scene-Driven Outputs Under Heavy Occlusion:** In scenarios of substantial occlusion, outputs should reflect the broader scene context rather than the occluded object, especially if the object does not align well with the scene's semantics.
  3. **Graded Relatedness Effects:** As object-scene relatedness decreases, model performance should systematically decline; objects that are highly congruent with the scene will be easier to identify and describe compared to those with lower congruence.
  4. **Attention Shifts:** When the target is heavily distorted or noisy, VLMs should redirect focus from the target region to contextual elements, paralleling human reliance on scene-based expectations.

2.2 Methodology

2.2.1 Dataset Development: Common Objects Out-of-Context (COOCO)

A key contribution is the COOCO dataset, extending COCO-Search18 [8] to systematically evaluate how VLMs handle scene-object semantic relationships. We introduce controlled variations in object–scene relatedness:

- **Original Images:** The target-present images from COCO-Search18, each featuring one instance of 18 possible object categories.
- **Generated Scenes:** An inpainting pipeline replaces the target with objects of medium or low semantic relatedness, quantified via ConceptNet Numberbatch [54].

All generated images are checked for realism. The final dataset spans thousand of images of diverse scenes (e.g., kitchens, streets, offices), with objects labeled as low, medium, or high in relatedness. A curated subset of manually filtered images is also provided.

2.2.2 Model Selection

We evaluate five leading VLMs trained or fine-tuned for grounding tasks: **Kosmos-2** [44], **Molmo** [10], **xGen-MM (BLIP-3)** [67], **Qwen2-VL** [63], and **LLaVA-OneVision** [29]. Each model performs referring expression generation on COOCO, producing text to identify a target object in images under various noise and relatedness conditions.

2.2.3 Experimental Design and Evaluation Metrics

Conditions

- **Noise Levels:** Gaussian noise or occlusion is applied at none, medium, or high levels to examine how models handle varying degrees of visual distortion.
- **Relatedness:** Objects with high, medium, or low semantic congruence with the scene are introduced to gauge the model's capacity for context-driven predictions versus susceptibility to semantic conflicts.
- **Noise Area:** Noise is introduced in the target region, the surrounding context, or both, allowing us to isolate the contribution of context from direct visual cues to the target.

Evaluation Metrics:

1. **RefCLIPScore:** A CLIP-based visual-text similarity measure [20] that gauges how well the generated text aligns with the target.
2. **Text-based Semantic Similarity:** Quantifies how accurately the generated expression aligns with the target label.
3. **Accuracy:** Assesses whether the model correctly identifies the target object.

Attention Deployment Analyses: We will also analyze attention maps in LLaVA-OneVision's encoder layers to understand how context informs internal representations.

Chapter 3

Dataset

Understanding the nuanced interplay between objects and their surrounding scenes is a critical challenge in cognitive science [61] and vision and language computational modeling [24]. Existing datasets have made significant strides in advancing object recognition, scene understanding, and visual attention modeling. However, they often fall short when it comes to capturing the complexity of semantic violations, scenarios where objects defy contextual expectations within a scene. This limitation hinders the ability to study how humans and machines process such incongruities and restricts progress in modeling more sophisticated cognitive behaviours in artificial systems.

With the goal of bridging the gap between psychology and AI research and fostering comparative studies on human visual attention and multimodal vision models, we introduce the *Common Objects Out-of-Context (COOCO)* dataset. A set of *COOCO* images is presented in Figure 3.1. This dataset is designed to provide a collection of images in which the semantic coherence of a scene is disrupted by the presence of an object with low or moderate semantic relatedness to the scene type. In each image, we identify an existing object that is semantically related to the scene, "clean" the image by removing it, and then replace it with a less semantically related object using image generative models. *COOCO* builds upon a framework previously established in visual perception psychology with the SCEGRAM dataset [39], scaling it up to meet the needs of research in vision and language modelling. It is meticulously curated to also support studies in visual perception psychology, particularly in the context of visual search tasks.

The proposed dataset includes 1,862 images sourced from COCO-Search18 [8], alongside their "clean" counterparts where target objects have been removed. Additionally, it contains 5,572 generated images featuring objects with medium semantic relatedness to the scene and 5,480 images with objects of low relatedness. Each original image is associated with a maximum of six generated variations, with up to three per relatedness level. In total, the dataset comprises 14,776 images.

3.1 Why a New Dataset?

This chapter explores the gaps in current datasets and establishes the motivation for creating a new dataset designed explicitly to address these limitations. By analyzing prominent datasets across domains such as computer vision, visual attention modeling and the few existing examples of datasets focusing on semantic scene violations, we highlight the challenges they face in annotating, representing, and scaling scenarios involving semantic incon-

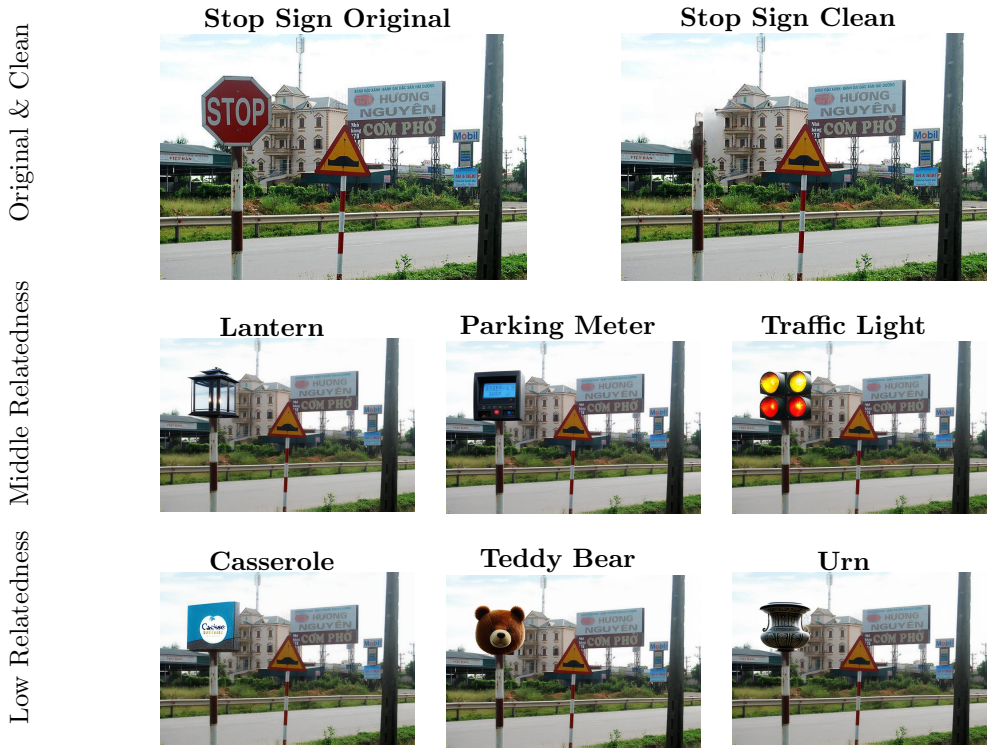


Figure 3.1: Visualization of a set of *COOCO* images from the "street" scene category, grouped by relatedness levels: The first row displays original and clean images, the second row shows images with medium relatedness scores, and the third row includes images with low relatedness scores.

gruities. Building on this foundation, we propose a new dataset that bridges these gaps, enabling deeper investigations into scene-object semantic relatedness and the development of multimodal models capable of handling such complexities.

The subsequent sections review key datasets and their limitations, culminating in a rationale for the creation of a dataset that balances realism, control, and diversity while facilitating advanced analysis of semantic violations in visual scenes.

3.1.1 Computer Vision Datasets

Microsoft COCO

The Microsoft COCO dataset [31] is widely recognized for its high-quality annotations of object localization and segmentation, making it a foundational resource for object recognition tasks. Furthermore, the inclusion of five written captions for each image has made COCO a cornerstone for vision-language modeling [47, 30]. However, its primary focus on commonly recognizable objects limits its ability to explore rare or unexpected object-scene relationships. Moreover, the dataset does not explicitly annotate or address semantic violations, making it unsuitable for studying contextual incongruities.

Visual Genome

The Visual Genome dataset [26] offers dense annotations of objects, attributes, and relationships within images, leveraging WordNet synsets for standardization. With over 42 region-based descriptions per image and an average of 17 question-answer pairs, it serves as a rich resource for modeling complex interactions between vision and language. These features make it particularly useful for tasks like visual question answering (VQA) and scene understanding. However, Visual Genome does not provide explicit annotations or mechanisms to evaluate object-scene relatedness or semantic violations within a scene, which limits its applicability for studying contextual incongruities.

SUN Database

The SUN database [66, 65] is notable for its extensive coverage of over 397 scene categories, making it a valuable resource for benchmarking scene recognition algorithms. However, its focus on scene-level categorization comes at the expense of detailed object-level annotations or explicit modeling of semantic relationships. This emphasis on broad scene recognition renders the dataset less suitable for examining contextual incongruities or object-scene semantic violations.

3.1.2 Visual Attention Modeling Datasets

SALICON

The SALICON dataset [23] captures human attention data during natural image exploration, focusing on saliency modeling. While it provides valuable insights into bottom-up attention patterns, it does not address top-down processes related to the semantic or contextual congruence of objects within scenes.

DIDEC

The DIDEC dataset [37] combines eye-tracking data with spoken image descriptions, offering insights into the interaction between visual attention and language production. However, it lacks explicit manipulations or annotations of semantic violations, focusing more on naturalistic viewing and description tasks rather than incongruities or contextual mismatches.

COCO-Search18 and COCO-FreeView

These datasets investigate attention behaviours under both goal-directed and free-viewing conditions. COCO-Search18 [8] explores search behaviour in target-present and target-absent scenarios, while COCO-FreeView [7] captures free-viewing patterns. Although valuable for understanding attention control, they do not address object-scene congruence or semantic violations.

Microwave-Clock Search Dataset

The Microwave-Clock Search Dataset [71] examines goal-directed search behaviour through controlled experiments focused on specific object categories (only analog clocks and microwaves). While it provides insights into search difficulty and object saliency, its narrow focus on a limited set of objects makes it unsuitable for studying broader semantic scene-object relationships.

3.1.3 Semantic Scene Violation Datasets

ObjAct Dataset

The ObjAct dataset [52] explores object congruency within human-action contexts, providing controlled real-life scenes with both congruent and incongruent object placements (e.g. a man drinking from a can/a man drinking from a potato). While it is effective for studying semantic violations in action-based scenarios, it does not address broader scene-object relationships unrelated to actions. Moreover, having been handcrafted, its resource-intensive creation process limits scalability.

Event Task Stimulus Set

The Event Task Stimulus Set [15] comprises photographs depicting people performing simple actions in ecologically valid contexts of varying complexity. Half of the images represent plausible events, while the other half depict implausible scenarios violating "world knowledge". The stimuli were validated by a group of 10 individuals for plausibility and balanced for attributes such as gender, age, and number of people depicted.

The dataset is particularly valuable for studying semantic violations related to human actions within realistic settings. However, its exclusive focus on action-based scenarios limits its applicability to broader studies of scene-object relationships or non-human semantic incongruities. Additionally, the reliance on subjective plausibility ratings may constrain the dataset's scalability and use in computational models.

Out-of-Context Dataset (OCD)

The Out-of-Context Dataset (OCD) [4] consists of 15,773 images spanning 36 object classes and six contextual conditions, generated using the VirtualHome environment in Unity across seven apartments with five furnished room types. The dataset explores normal context (images with objects in typical locations) and no-context conditions (images with surrounding pixels replaced by noise or uniform grey). Additional conditions include gravity (images with objects lifted off their support), object co-occurrences (images placing objects in atypical locations based on human judgments), a combined gravity and object co-occurrence condition (images placing lifted objects in unlikely locations such as walls or doorways), and size manipulations (images where objects are resized to 2-4 times their original size). The dataset ensures target objects remain centered, avoiding occlusions and collisions, to study object placement and contextual understanding. However, a limitation of the OCD is that, being generated in a virtual environment, it lacks the realism of real-world images. The synthetic nature of the dataset may not fully capture the complexity of real-world textures, lighting conditions, and object interactions, potentially limiting its generalizability to real-world scenarios.

Cut-and-Paste Dataset

The Cut-and-Paste dataset [70] is a synthetically generated training dataset based on the cut-and-paste method, which combines real object images with diverse background scenes to create composite images for object detection tasks. The dataset is constructed using two distinct source domains: (1) a foreground source domain consisting of object instances captured in a controlled environment with uniform lighting and a neutral background, and (2) a background source domain derived from publicly available scene datasets that provide a wide variety of environmental contexts. Object masks are either manually annotated or

generated using automated segmentation techniques, enabling precise extraction of object regions from their original images. These segmented objects are then placed onto randomly selected background images to synthesize realistic training samples. However, this standard cut-and-paste approach introduces an unbalanced domain gap due to the inherent differences between the foreground and background source domains, which can negatively impact model generalization. To mitigate this issue, the dataset incorporates background simplification and foreground diversification. Background simplification is achieved by applying image processing techniques such as Gaussian blurring, grayscale conversion, and color quantization, thereby reducing the complexity and diversity of background images. Foreground diversification is implemented using a generative adversarial network (GAN)-based model, which introduces stylistic variations to object images, expanding the foreground domain and improving domain adaptation.

A drawback of cutting-and-pasting is the introduction of artifacts such as unnatural lighting, object boundaries, sizes and positions.

SCEGRAM Dataset

The SCEne GRAMmar manipulations (SCEGRAM) dataset [39], that is the dataset we took as an example for building *COOCO*, was specifically designed to investigate semantic and syntactic violations in scenes, offering a controlled framework for studying congruent and incongruent object placements. The dataset consists of 744 scene images across six key conditions:

- **Consistent control condition (CON)**: A semantically and syntactically consistent object is placed in a probable location (e.g., toilet paper on a toilet paper holder).
- **Inconsistent-semantics condition (SEM)**: A semantically incongruent object is placed in a syntactically consistent location (e.g., a cup on a toilet paper holder).
- **Mild inconsistent-syntax condition (SYN)**: A semantically congruent object is placed in a physically possible but syntactically incongruent location (e.g., toilet paper on a toilet seat cover).
- **Mild double-inconsistency condition (SEMSYN)**: A semantically incongruent object is placed in a syntactically incongruent but physically possible location (e.g., a cup on a toilet seat cover).
- **Extreme inconsistent-syntax condition (EXSYN)**: A semantically congruent object is placed in a syntactically inconsistent and physically impossible location (e.g., toilet paper hovering midair above a toilet).
- **Extreme double-inconsistency condition (EXSEMSYN)**: A semantically incongruent object is placed in a syntactically inconsistent and physically impossible location (e.g., a cup hovering midair above a toilet).

A representative sample of each condition is illustrated in Figure 3.2. Each scene was photographed with and without the critical objects (present and absent conditions) to minimize lighting and environmental changes. Object-only images, photographed against a uniform white background, allow for further testing of object recognition and priming effects. Areas of interest (AOIs) were manually annotated to support tasks such as eye-tracking and saliency validation.

While SCEGRAM provides a robust framework for studying semantic violations, it is limited in scope, featuring only 62 scenes and a small set of objects. The reliance on paired congruent and incongruent objects within each scene restricts its diversity and scalability for computational research. A larger, more diverse dataset is required to facilitate a deeper exploration of semantic and syntactic incongruities in visual scenes.

Figure 3.2: Example images from the SCEGRAM dataset, illustrating six conditions of semantic and syntactic (in)congruence in object placements within scenes. Source: [39]

3.1.4 Limitations of Existing Datasets

In our evaluation of existing datasets, we identified several critical shortcomings that compromise their suitability for our research objectives:

- **Lack of Explicit Semantic Violation Annotations:** Mainstream datasets (e.g., Microsoft COCO, Visual Genome, SUN Database) focus on object recognition and scene categorization but do not explicitly label or address semantic incongruities.
- **Emphasis on Common Objects and Typical Scenes:** These datasets primarily capture common objects and standard scene compositions, limiting the exploration of rare or unexpected object-scene relationships.

- **Narrow Focus in Visual Attention Datasets:** Datasets such as SALICON, COCO-Search18, and DIDEC concentrate on saliency and goal-directed tasks without considering the impact of semantic congruence on attention.
- **Limited Scope and Scalability in Semantic Violation Datasets:** Datasets addressing semantic violations (e.g., ObjAct, Event Task Stimulus Set, SCEGRAM) are often confined to specific contexts (like human actions), feature small sample sizes, or rely on subjective ratings, hindering broader applicability.
- **Challenges with Synthetic Data:** Synthetic datasets (e.g., Out-of-Context Dataset, Cut-and-Paste Dataset) face issues like reduced realism, artifacts, and domain gaps between foreground and background elements.

3.1.5 Need for a Dataset Emphasizing Scene-Object Semantic Relatedness with Violations

To address these limitations, there is a pressing need for a large-scale dataset that explicitly emphasizes scene-object semantic relatedness while incorporating semantic violations. Such a dataset should capture the complexity of contextual incongruities, allowing for a deeper exploration of how humans and artificial systems process these violations. It must balance realism, control, and diversity, offering a wide range of objects, scenes, and violation types.

Key features of this dataset would include explicit annotations for semantic congruence and incongruence, scalability to support computational modeling, and versatility to accommodate various research domains, from cognitive science to vision-language modeling.

By bridging the gaps in existing resources, such a dataset would open new avenues for studying contextual incongruities, enabling advances in cognitive modeling and artificial systems' understanding of complex visual semantics.

3.2 Dataset Construction Process

3.2.1 Initial Dataset Selection

The COCO-Search18 dataset [8] was selected as the foundation for developing a new image dataset due to its suitability for tasks involving visual search. Built upon the Microsoft COCO dataset [31], COCO-Search18 offers a curated subset of COCO's images, specifically annotated with human gaze fixation data from visual search experiments, where participants were given a target object and they searched for it in the image. This makes it uniquely valuable for tasks requiring insights into human attention during object-focused searches.

The idea of building our *COOCO* dataset emerged from a comprehensive research initiative examining human attention across diverse scenarios, including complex referential and linguistic tasks. Within this wider framework, COCO-Search18 serves a more specific role: it provides the groundwork for a focused study on Referring Expression Generation (REG). COCO-Search18's curated gaze data and visual search setup offer a unique opportunity to align computational models with human cognition for this purpose.

COCO-Search18 was designed with strict criteria to ensure its usability for visual search and attention modeling:

1. **Exclusion of images with people or animals**: To minimize biases arising from natural tendencies to fixate on these elements, ensuring target-driven search behaviors are studied.

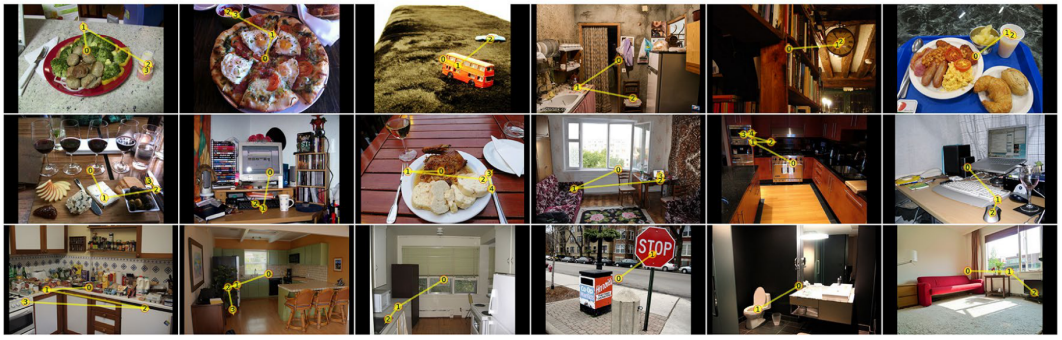


Figure 3.3: Examples of target-present images for each of the 18 target categories. Yellow lines and numbered discs indicate a representative search scanpath from a single participant. Source: [8]

2. **Single-instance targets**: Only images containing one instance of the target were included to eliminate ambiguity during search tasks.
3. **Target size constraints**: Targets were required to occupy between 1% and 10% of the image area to balance task difficulty.
4. **Target location**: Centrally positioned targets were excluded to avoid interference from the pre-positioned central gaze at the start of each trial.
5. **Aspect ratio restrictions**: Images with a width/height ratio outside the range of 1.2–2.0 were excluded to maintain natural viewing conditions.
6. **Minimum category size**: Object categories with fewer than 100 images meeting these criteria were excluded to ensure sufficient data for training AI models.

Additional refinements included excluding occluded or difficult-to-recognize targets using object detectors and manual reviews to remove objectionable content. These steps resulted in a final dataset of 3,101 images containing the target object, spanning 18 object categories: *bottle, bowl, car, chair, (analogue) clock, cup, fork, keyboard, knife, laptop, microwave, (computer) mouse, oven, potted plant, sink, stop sign, toilet, and TV*. Additionally, the dataset included other 3,101 images that did not contain any target, for a total number of 6,202 images. An example of COCO-Search18 images is shown in Figure 3.3.

COCO-Search18 is further complemented by the COCO-FreeView dataset, which includes free-viewing data collected as part of the same effort [7]. This dataset uses the same natural images as COCO-Search18 but is annotated with human eye fixations recorded during a free-viewing task.

Of the datasets mentioned, I directly accessed the publicly available subset for the creation of my new dataset. Specifically, 2,241 images contained the target object (present), and 2,094 images did not contain the target (absent), for a total of 4,335 images. Only the target present images are used to generate the final dataset.

3.2.2 Modeling Scene-Object Semantic Relatedness

To achieve our goal of generating a dataset featuring images with semantic violations, we built upon the annotations available in the COCO-Search18 dataset. However, additional

information was needed to measure semantic relatedness between scenes and target objects and to define semantic violations of scene grammar quantitatively.

Scene Classification

The first step was to identify the scene depicted in each image. To achieve this, we classified each image into a specific scene type. For this classification task, we fine-tuned a Vision Transformer (ViT) model [14, 64] using the Scene UNderstanding (SUN-397) dataset [65, 66].

The Vision Transformer (ViT) is a transformer encoder model, conceptually similar to BERT [12], pre-trained in a supervised manner on the large-scale ImageNet-21k dataset [48], which contains 21,000 classes and images at a resolution of 224×224 pixels. Following pretraining, the model was fine-tuned on ImageNet [11], a dataset with 1 million images across 1,000 classes, also at 224×224 resolution. In ViT, input images are divided into fixed-size patches (16×16 pixels), which are linearly embedded into a sequence of vectors. A [CLS] token is appended at the start of this sequence to be used for classification tasks. Absolute positional embeddings are added to the patch embeddings before the sequence is processed by the Transformer encoder. This pretraining allows the model to learn rich feature representations of images that can be used for various downstream tasks [14].

A linear layer is placed on top of the pre-trained encoder to train the standard classifier. This linear layer is applied to the [CLS] token, as the last hidden state of this token represents the entire image.

The SUN-397 dataset is a curated subset of the larger SUN dataset, which consists of 899 categories and 130,519 images. This subset includes 397 categories, specifically selected because each contains at least 100 unique photographs, resulting in a total of 108,753 images [66]. The dataset was split into training and validation sets with a ratio of 90 : 10.

We fine-tuned the Vision Transformer (ViT) model for 15 epochs using a learning rate of 2×10^{-4}, a weight decay of 0.1, and a batch size of 64 for both training and evaluation. Metrics were logged, and performance was assessed at the end of each epoch. The default cross-entropy loss function was used, and the model achieved a validation accuracy of 79%.

After an initial classification of the Coco-Search18 images, we downsampled the categories to 25 carefully balancing high-frequency occurrences with semantic diversity and then reclassified the images using these categories. This minimizes the risk of both underrepresentation of certain categories and excessive variation, ensuring that the final dataset remains both robust and meaningful.

Set of Objects and Size Data

To determine which objects would replace the target objects in the final images, we needed additional information about the candidate objects. For this, we relied on the THINGSplus dataset [55].

THINGSplus is an extension of the THINGS database [18], designed to support research across psychology, neuroscience, and computer science. It provides enriched norms and metadata for 1,854 systematically sampled object concepts and their 26,107 high-quality, naturalistic images. The dataset includes 53 superordinate categories, typicality ratings for all category members, and concept-specific norms for properties such as real-world size, manmadeness, preciousness, liveliness, and graspability, among others [55].

We began the filtering process by selecting a subset of object names based on specific criteria. Initially, objects with a typicality score between 0.3 and 1 were included, to exclude highly atypical category members. Next, animate objects from categories such as animals,

body parts, farm animals, insects, mammals, sea animals, and seafood were excluded. Objects belonging to the women's clothing category were also excluded, as the image generative model used to generate the final versions of the images in *COOCO* (see Section 3.2.3), tended to produce body parts or full human figures alongside these objects. This was done to keep the final dataset in conformity with the first criteria the creators of COCO-Search18 used, namely, the exclusion of images with people or animals, to avoid biases during visual search.

To ensure a realistic effect of the final object replacement we used the real-size data from the THINGSplus dataset to select every time an object with a comparable size with respect to the target object.

Measuring Scene-Object Semantic Relatedness

To measure the scene-object semantic relatedness we decided to use the cosine distance between the ConceptNet Numberbatch embeddings [54] of their scene and object labels.

ConceptNet Numberbatch uses an ensemble approach combining the semantic vectors from Word2vec [36] and GloVe [45], which learn how words are associated with each other from large text corpora, with ConceptNet, a knowledge graph that draws on expert-created resources and crowdsourced knowledge [54].

We chose this set of embeddings based on insights from a recent study [17], which explored the relationship between object semantics and attention by integrating a vector-space model of semantics with eye-tracking data in scene analysis. In this approach, ConceptNet Numberbatch embeddings were used to construct a concept map that indexed the spatial distribution of semantic similarity among objects within a scene. The study demonstrated a strong positive correlation between the semantic similarity of a scene region and the focus of viewers' attention, with greater attention directed toward regions with higher semantic relatedness to the scene. These findings underscore the critical role of object semantics in guiding visual attention in real-world scenes and directly validate that ConceptNet Numberbatch embeddings effectively model object semantics within the visual domain.

3.2.3 Image Generation Pipeline

Figure 3.4 illustrates the pipeline used to clean and replace the target object in COCO-18 images with a semantically unrelated candidate, visually outlining the process.

Image Preparation

The image generation pipeline processes each image of the COCO-Search18 dataset through a series of steps to replace the target object within it with an object that is not related with the scene and thus semantically violates the scene grammar. First, the pipeline retrieves the relevant data for each image, including the target object, its bounding box, the scene category, and a white mask on a black background representing the target object. This mask is created using the COCO object segmentation annotations, expanded by 20% of the original area, to be used with the inpainting model for object removal. This step is common in object removal workflows with image generative models, as it enhances the removal quality by better capturing the target region.

Next, the target object is removed from the image by feeding the mask to the Large Mask Inpainting (LaMa) model [56]. Afterward, a square-padded version of the cleaned image is generated, along with a corresponding rectangular mask of the target object's area. These two components are then prepared as inputs for the generation model.

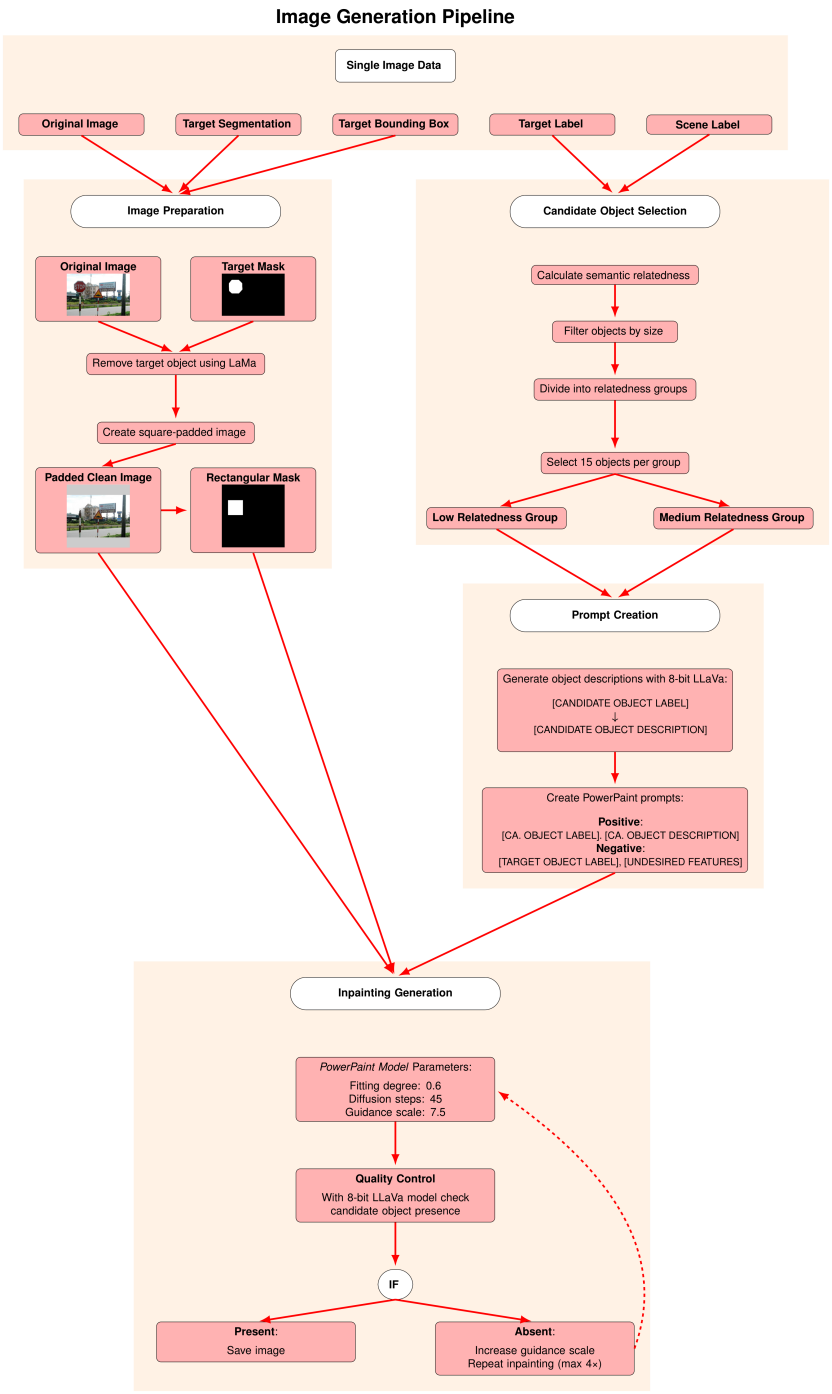


Figure 3.4: Illustration of the pipeline.

Candidate Object Selection

The candidate objects for replacement are selected based on their semantic relatedness to the scene and their size similarity to the target object. First, the semantic relatedness of each object to the scene is computed using cosine similarity between the scene vector and the object vectors. Objects with semantic relatedness close to 0 are considered less related to the scene, while those with relatedness close to 1 are more contextually appropriate. Simultaneously, the real-size data from the ThingsPlus dataset [55] of each object is compared to that of the target object, with the size difference being used as a score. Objects with a size difference greater than a specified threshold (25 units) are filtered out. The remaining objects are then categorized into two groups based on their semantic relatedness: low (close to cosine similarity value at 0) and medium (close to 0.25). Then the 15 objects closest to the score are selected from each group.

Prompt Creation

To provide a richer textual context for the inpainting model and enhance generation quality, detailed descriptions of candidate objects are produced. A single candidate is randomly selected, and a prompt is constructed to pass it to an 8-bit quantized [22] version of the LLaVa model [33]. The goal is to produce a more descriptive account of the object's appearance. The following prompt is used:

```
"Write a general description of the object [ARTICLE][CANDIDATE OBJECT
↪  LABEL]. Focus only on its appearance. Be concise."
```

Here, `[ARTICLE]` corresponds to the appropriate article based on the initial letter of `[CANDIDATE OBJECT LABEL]`, which is the name of the selected candidate object.

The description generated by LLaVa is then appended to the object label, forming a more detailed prompt for the inpainting model. This prompt is designed to generate the candidate object as a replacement for the target object in the image. An example of the final prompt is as follows:

```
"[ARTICLE][CANDIDATE OBJECT LABEL]. [CANDIDATE OBJECT DESCRIPTION]"
```

To further refine the output, a negative prompt is provided to the inpainting model, guiding the generation process away from undesired features. Below is an example of a negative prompt:

```
"[TARGET OBJECT LABEL], humans, people, person, body, face, head, hands,
↪  legs, arms, torso, skin, eyes, mouth, fingers, feet, hair, human-like
↪  figures, silhouettes, limbs, human anatomy, human features, mannequins,
↪  dolls, humanoid shapes"
```

Inpainting Generation

Using the positive and negative prompts, a shape-guided inpainting task is initiated to replace the target object in the image with the candidate object. Image inpainting involves filling specified regions in an image with visually plausible content. With the rise of text-to-image (T2I) models, inpainting has gained prominence as a flexible and interactive method for refining generated images by masking and regenerating unsatisfactory regions [76].

To achieve our objective of replacing the target object with the candidate object, we employed the inpainting model PowerPaint [76]. PowerPaint is a versatile model that excels

in both text-guided object synthesis and context-aware image filling. It is built upon the pre-trained Stable Diffusion model [49] and it utilizes specialized learnable task prompts and tailored training strategies to handle diverse inpainting tasks within a single model [76].

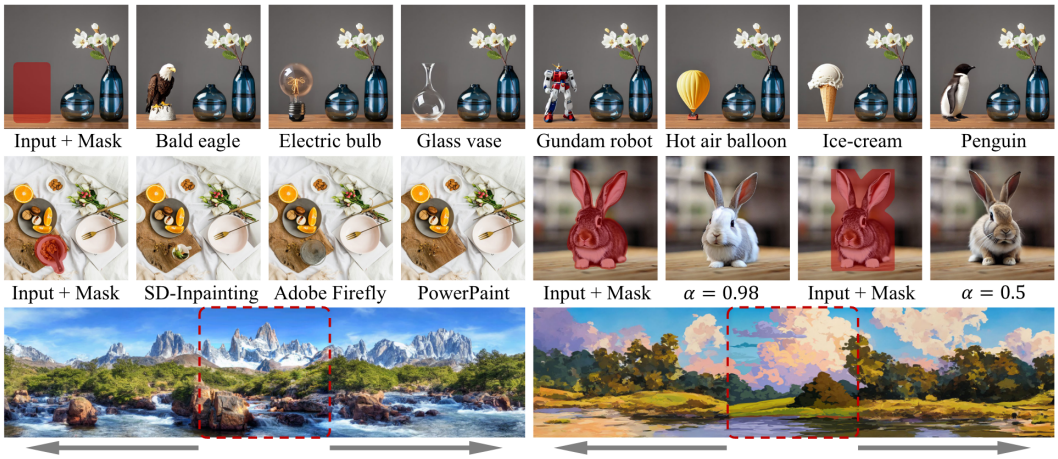


Figure 3.5: Examples of various inpainting tasks supported by PowerPaint, including text-guided object inpainting, object removal, shape-guided object inpainting with controllable shapefitting, outpainting, and more. Source: [49]

The inpainting process requires specifying three key parameters playing a critical role in shaping the final output.:

- The fitting degree (ranging from 0 to 1), determines how closely the generated objects align with the shape of the masks. To encourage object generation that is not overly constrained by mask shapes, we use the recommended value of 0.6 for this parameter.
- The number of diffusion steps (0 to 50), primarily affects the quality of the generated image but comes with a trade-off in computational cost. We retain the default value of 45, as specified in the PowerPaint model.
- The guidance scale factor (0 to 30), controls the extent to which the generation adheres to the provided text input. Initially, we set this parameter to 7.5.

Quality Control Over Generation

To ensure the generated image contains the candidate object, an image quality filter is implemented using the 8-bit quantized LLaVa model. The generated image and prompt are evaluated using the following query:

```
"[INST] <image> Is there [ARTICLE] [CANDIDATE OBJECT LABEL] in the image?
↪  [CANDIDATE OBJECT DESCRIPTION]. Answer only with 'Yes' or 'No'.
↪  [/INST]"
```

If the model responds "Yes," the image is saved. If "No," the guidance scale factor is incremented by 7.5, and the generation process is repeated. The process is limited to 4 attempts per candidate object-image pair. If all attempts fail, the candidate object is discarded, and another is sampled. For each image, three candidate objects are selected from both low and middle relatedness groups, generating six images per input.

Final Automatic Filtering

After completing the generation process, the entire dataset underwent an additional round of filtering using the full-precision version of the LLaVa model. While the generation phase relied on the 8-bit quantized model to economise on hardware usage during the resource-intensive data generation phase, the full-capacity model was employed during this step to ensure the final dataset met the highest quality standards. The same prompt used during quality control in the generation phase was applied here. However, in this stage, if the model's response was negative, the corresponding image was marked as filtered out in the final dataset.

3.2.4 Manual Filtering

To create a smaller, high-quality subset of the dataset suitable for visual attention experiments with human participants, we applied a final round of manual curation. This process specifically focused on the low-relatedness group of images and their corresponding original versions. These images were split in three sets among three researchers.

Images were excluded if the object was completely unrecognizable, disrupted the scene's perspective (e.g., objects appearing outside a window in a scene viewed from inside), or lacked a cohesive scene structure (e.g., grids of unrelated objects). Objects replaced with the same type but labeled as "low similarity" (e.g., a "chair" replaced with a "rocking chair") or partially replaced (e.g., only the face of a clock) were also excluded. We also discarded blurry, distorted, or low-quality generated objects, as well as black-and-white images where the replacement object appeared in color. Scenes that were misclassified by type or contained poorly labeled objects were removed. Frequently misgenerated objects, such as elongated items (e.g., swords, crutches), "breakfast," or human figures, were excluded due to their low recognizability. Generated images featuring altered regions outside the bounding box or showing extraneous, unintended changes were also removed. Only high-quality images presenting complete and coherent scenes were retained. If all the generated images for a given original image were excluded, the original image was also removed from the dataset. The final curated dataset consists of 1,916 low-relatedness images and 953 original images, for a total of 2,870 images.

Chapter 4

Models

To perform a Reference Expression Generation (REG) task with large pre-trained Vision-Language Models (VLMs), we required them to interpret region-of-interest (ROI) specifications directly from input coordinates. This capability enables the models to refer to specific areas of an image without relying on a textual description of the region's content. To ensure this, we selected models that had been exposed during pre-training or fine-tuning to tasks that develop grounding abilities. Recent definitions describe these abilities as the precise mapping of specific spatial regions in an image to corresponding linguistic elements, thereby enabling direct object referencing and enhancing both contextual and referential precision [44].

4.1 Grounding Abilities

Grounding ability is a crucial aspect of visual-language models (VLMs) that enhances human-AI interaction in vision-language tasks. It allows users to reference objects or regions within an image by pointing, rather than relying solely on textual descriptions. This capability enables the model to understand and associate spatial locations with language, facilitating more accurate and context-aware responses. Furthermore, grounding ability supports visual answers, such as bounding boxes, which enhance referential precision and resolve coreference ambiguities compared to text-only responses [44].

By leveraging grounding, models can link noun phrases and referring expressions in generated text responses to specific image regions, producing more comprehensive and informative answers. This is achieved by encoding spatial coordinates as location tokens or regular text tokens integrated into the text sequence, effectively creating a "hyperlink" between the visual and linguistic modalities. As a result, these models exhibit strong performance on grounding tasks such as phrase grounding and referring expression comprehension, as well as referring tasks like REG [44].

Recent advances in grounding have further expanded its application in multimodal AI. For example, the PixMo-Points dataset introduced pointing data that enables models to answer questions by directly pointing to image regions, improving both naturalness and accuracy in tasks such as object identification and counting [10]. An example of this is provided in Figure 4.1. Referential Dialogue (RD) extends this concept by allowing users to engage in conversations while referencing precise image regions, making multimodal AI interactions more intuitive [6].

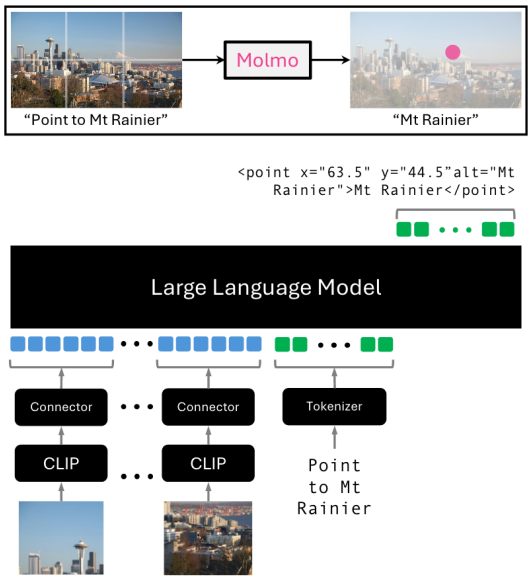


Figure 4.1: Molmo architecture and a sample of PixMo data used for model training. Source: [10]

Grounding ability is particularly relevant for vision-language positioning tasks, including Visual Question Answering (VQA), Referring Expression Comprehension (REC), and Described Object Detection. Traditional methods for encoding position inputs include concatenating cropped image patches, using binary masks, or employing learned positional encodings. However, newer approaches, such as GPT4ROI, incorporate spatial instruction tuning by interleaving region-of-interest features with textual embeddings, enhancing fine-grained visual understanding [74]. An illustration showing this process is shown in figure 4.2. Similarly, LLaVA-Grounding integrates grounding with visual chat, enabling models to maintain conversational coherence while associating text-based references with specific image regions [73].

After an initial survey of plausible models possessing visual grounding abilities, we selected a subset based on their ease of implementation. In the following sections, we provide a concise overview of the chosen models, detailing their input format requirements for spatial information. Depending on the model architecture, spatial inputs could take the form of bounding boxes, point coordinates, or location tokens. These elements were either integrated within the text input sequence or appended as a separate representation in the final input structure.

4.2 Kosmos-2

KOSMOS-2 [44] is a grounded multimodal large language model that extends the capabilities of its predecessor, KOSMOS-1 [21], by incorporating grounding and referring functionalities. The model follows a Transformer-based causal language model architecture, leveraging next-token prediction as its primary training objective. It processes various input modalities, including text, images, and interleaved image-text pairs, while introducing a novel mechanism to establish direct associations between textual spans and specific image regions.

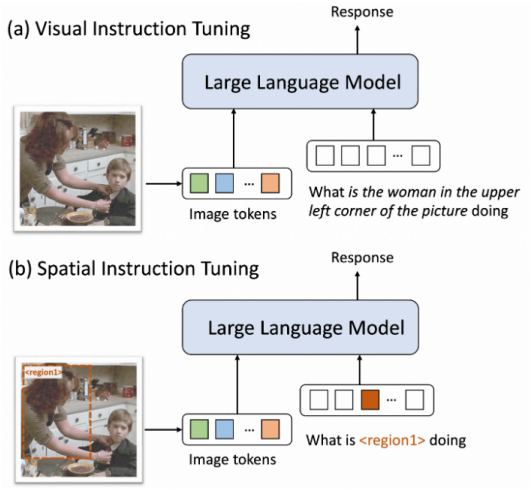


Figure 4.2: Comparison of visual instruction tuning on image-text pairs and spatial instruction tuning on region-text pairs. Source: [74]

The model maintains the architectural foundation of KOSMOS-1, comprising a vision encoder, a resampler module, and a language model. The vision encoder is responsible for extracting visual features from the input images, which are subsequently processed by the resampler module to generate compact image embeddings. These embeddings, along with textual tokens and spatially grounded information, form the input representation for the Transformer-based backbone.

KOSMOS-2 is trained on an expanded dataset, incorporating grounded image-text pairs in addition to the multimodal corpora used in KOSMOS-1. The training loss is defined over discrete tokens, including both textual and location tokens, enabling the model to learn precise mappings between image regions and corresponding text spans. This training paradigm equips KOSMOS-2 with the ability to not only generate textual descriptions of images but also predict and interpret bounding boxes, thereby facilitating grounded visual reasoning.

This model's version we used is specifically the Hugging Face version:

- `microsoft/kosmos-2-patch14-224`

4.2.1 Encoding Spatial Information

KOSMOS-2 introduces a systematic approach for processing and integrating spatial information into the model's input representation. Specifically, the model employs a discretization strategy to convert continuous bounding box coordinates into discrete location tokens, which are seamlessly integrated with textual tokens in a unified manner. To achieve this, the model first defines a spatial grid over the image by dividing its width W and height H into $P \times P$ segments, where each segment (or bin) corresponds to $(W/P) \times (H/P)$ pixels. Each bin is assigned a unique location token representing its coordinates, effectively transforming the spatial information into a discrete vocabulary that can be processed by the language model. A bounding box in an image, characterized by its top-left corner (x_1, y_1) and bottom-right corner (x_2, y_2), is encoded using the corresponding location tokens of these points. This representation is structured as:

$$\langle box \rangle \langle loc_1 \rangle \langle loc_2 \rangle \langle /box \rangle$$

where $\langle box \rangle$ and $\langle /box \rangle$ denote the boundary tokens marking the presence of a bounding box. If a textual span is associated with multiple bounding boxes, the location tokens of each bounding box are concatenated using a delimiter token:

$$\langle box \rangle \langle loc_{i1} \rangle \langle loc_{i2} \rangle \langle delim \rangle \dots \langle loc_{j1} \rangle \langle loc_{j2} \rangle \langle /box \rangle$$

These location tokens are mapped to embeddings via a lookup table and processed jointly with the textual tokens. The final input sequence follows a structured format inspired by hyperlink annotations, explicitly linking textual spans to visual regions. For instance, an input representation incorporating both text and grounded bounding boxes may be formatted as:

$\langle s \rangle \langle image \rangle$ Image Embedding $\langle /image \rangle$ $\langle grounding \rangle$
$\langle p \rangle$ Text Span $\langle /p \rangle$ $\langle box \rangle$ Location Tokens $\langle /box \rangle$ $\langle /s \rangle$

where $\langle grounding \rangle$ explicitly signals the requirement to associate textual descriptions with visual elements. Through this structured approach, KOSMOS-2 effectively learns to establish bidirectional mappings between textual entities and visual regions, supporting a range of downstream tasks such as grounded image-captioning, visual question answering, and referring expression comprehension and generation.

4.3 Molmo

Molmo [10] is a vision-language model (VLM) designed to process and generate descriptions of images. It follows a Transformer-based architecture, integrating a vision encoder and a language model. The model incorporates spatial grounding functionalities to improve visual understanding and multimodal reasoning.

The architecture consists of a vision encoder, a connector module, and a decoder-only Transformer LLM. The vision encoder extracts visual features from input images, which are then projected into the language model's input space through a Multi-Layer Perceptron (MLP). A pooling mechanism reduces computational costs while preserving crucial visual information. The language model is available in multiple variants, including fully open-weight models (OLMo-7B-1024, OLMoE-1B-7B) and open-weight Qwen2 models (7B and 72B parameters), offering trade-offs between efficiency and performance.

Molmo is trained on a diverse dataset incorporating dense captioning, document-focused question-answer pairs, analogue clock reading tasks, and datasets incorporating spatial grounding using 2D points.

This model's version we used is specifically the Hugging Face version:

- `cyan2k/molmo-7B-O-bnb-4bit`

4.3.1 Encoding Spatial Information

Molmo integrates spatial information using the PixMo-Points dataset, which introduces explicit point-based references within images. The dataset enables the model to:

- Identify and point to objects based on textual descriptions.
- Count objects by explicitly marking all instances.

- Use pointing as a visual explanation mechanism in question-answering tasks.

Human annotators provide point-based references in images, describing objects and marking absent entities to help the model learn to handle such cases. The dataset comprises 2.3 million question-point pairs from 428,000 images.

Specifically, object locations are represented in a fixed-scale coordinate system, where positions are mapped to a defined range of [0, 100]. For instance, an input query might take the form:

```
describe the object at point x=63, y=44
```

This structured approach supports object recognition, counting, and multimodal interactions.

4.4 xGen-MM (BLIP-3)

xGen-MM (BLIP-3) [67] is a vision-language model that integrates vision and language components for multimodal understanding. The architecture consists of a Vision Transformer (ViT), a vision token sampler based on the perceiver resampler, and a pre-trained Large Language Model (phi3-mini). It processes interleaved multimodal inputs, incorporating structured spatial grounding to improve object localization and referring expression comprehension.

The vision processing pipeline encodes images at high resolution, splitting them into multiple patches to retain fine-grained details. To manage computational costs, the perceiver resampler downsamples vision tokens, reducing sequence length while maintaining essential information.

This model's versions we used are specifically the Hugging Face versions:

- `Salesforce/xgen-mm-phi3-mini-instruct-singleimg-r-v1.5`
- `Salesforce/xgen-mm-phi3-mini-instruct-r-v1`

4.4.1 Encoding Spatial Information

xGen-MM integrates spatial information using a structured approach based on the BLIP3-GROUNDING-50M dataset, which annotates object locations extracted via state-of-the-art image tagging and object detection models. The spatial encoding follows three distinct formats:

1. Bounding box coordinates: `x1, y1, x2, y2`
2. Descriptive spatial relations: "starts at (x1, y1) and extends up to (x2, y2)"
3. Relative positioning: "top-left corner of the image"

During training, grounded spatial information is integrated into text captions, reinforcing the model's ability to align semantic concepts with visual features. This structured annotation supports fine-grained object localization and multimodal reasoning.

4.5 Qwen2-VL

Qwen2-VL [63] is a multimodal model available in multiple scales (2B, 7B, and 72B). It follows the Qwen-VL framework, integrating a 675M-parameter Vision Transformer (ViT) with the Qwen2 series of Large Language Models (LLMs). The model introduces dynamic resolution support, enabling it to process images of arbitrary resolutions efficiently.

The model employs 2D Rotary Position Embeddings (2D-RoPE) to handle variable-resolution images, replacing absolute position embeddings. A multi-layer perceptron (MLP) compresses adjacent visual tokens, optimizing memory consumption while maintaining crucial spatial information.

This model's version we used is specifically the Hugging Face version:

- `Qwen/Qwen2-VL-2B-Instruct-GPTQ-Int8`

4.5.1 Encoding Spatial Information

For representing locations, the model normalizes bounding box coordinates within a fixed range of [0,1000) and represents them as:

`(Xtop left, Ytop left), (Xbottom right, Ybottom right)`

These coordinates are enclosed within $< |box\_start| >$ and $< |box\_end| >$ tokens, explicitly linking textual descriptions to visual regions. Object references are defined using specialized tokens, enabling precise multimodal interactions.

4.6 LLaVA-OneVision

LLaVA-OneVision [29] is a multimodal model that integrates image features into a language model using a SigLIP-based [72] Vision Encoder, a two-layer MLP projector, and the Qwen-2 [68] Large Language Model. A comprehensive description of this model can be found in Section 7.1, where we conduct an attention deployment analysis on it.

This model's version we used is specifically the Hugging Face version:

- `llava-hf/llava-onevision-qwen2-0.5b-si-hf`

4.6.1 Encoding Spatial Information

LLaVA-OneVision [29] is trained on a broader and more diverse set of instruction-tuning data compared to its predecessors. While earlier models like LLaVA [33] primarily relied on single-image instruction data, LLaVA-OneVision incorporates a more comprehensive dataset, including single-image, multi-image, and video-based instructions. A defining feature of this training data is its reliance on bounding boxes, which provide essential object localization and spatial information. This structured representation enhances the model's capacity to interpret and reason about visual content across different scenarios, improving its generalization in multimodal tasks [33]. Additionally, LLaVA-OneVision has been extensively trained on Visual Grounding tasks, where the model outputs bounding boxes for specific targets in an image or generate region captions—short descriptions for specific areas within an image. Notably, for single-image training, datasets such as RefCOCO [69] (50,586 samples) and Visual Genome [26] (86,417 samples) have been used, while in the multi-image setting, datasets like WebQA [5] (9.3K samples) contribute to the model's ability to process and contextualize information across multiple visual inputs.

In instruction-tuning data, spatial information is encoded directly in the text input using bounding boxes in the format $[x_1, y_1, x_2, y_2]$, where coordinates are normalized within the [0,1] range.

Chapter 5

Experiment

This experiment is designed to address key research questions concerning how Vision-Language Models (VLMs) leverage contextual information for object recognition. In particular, it evaluates the extent to which these models rely on scene context when faced with varying degrees of semantic congruence between an object and its background, and how robustly they perform under different levels and configurations of visual noise. By systematically manipulating both the object-scene relatedness and the noise intensity/area, the experiment directly tests the conclusions of Juncker et al. [24] in a structured setting introducing also the semantic relatedness variable .

To achieve this, each model is provided with an image and a specified target region, defined by a bounding box, and is prompted to identify the object within that area using a structured text prompt. The experimental design incorporates multiple conditions: baseline images (with no noise), images with moderate noise, and images with high noise, applied either solely to the target, solely to the context, or to both regions. This setup enables us to isolate whether models depend more on the object's inherent visual features or on its surrounding context when recognition is challenged .

Furthermore, by introducing semantic violation conditions that vary the degree of relatedness between the object and its scene (high, medium, low), the experiment assesses the models' behaviour to contextual anomalies.

5.1 Prompt Design

To assess the model's performance, a structured text prompt is used. The prompt explicitly refers to the bounding box coordinates within the image and asks the model to identify the object in that region. The prompt follows this format:

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

This design ensures that the model's response is focused on object recognition within the specified region. The spatial information relative to the target is adjusted based on the input format required by each evaluated model.

5.2 Noise Injection

To assess the model's robustness, Gaussian noise is applied to images in different configurations. An image I is defined as:

$$I : \{0, \ldots, H-1\} \times \{0, \ldots, W-1\} \to \{0, \ldots, 255\}^C,$$

where H and W are the image dimensions, and C is the number of channels (e.g., $C = 3$ for RGB images). Given a bounding box $B = (x, y, w, h)$, where (x, y) is the top-left coordinate and w and h represents the width and height, the affected region R_B consists of all pixels within:

$$R_B = \{(i, j) \mid y \leq i < \min(y + h, H), \quad x \leq j < \min(x + w, W)\}.$$

Gaussian noise is sampled from the normal distribution:

$$N(i, j) \sim \mathcal{N}(0, (\sigma_{\max} \cdot \lambda)^2),$$

where $\sigma_{\max} = 255$ and $\lambda \in [0, 1]$ controls the noise intensity. The perturbed pixel values are computed as:

$$I'(i, j) = \text{clip}(I(i, j) + N(i, j), 0, 255),$$

where the clipping function ensures valid pixel values:

$$\text{clip}(v, 0, 255) = \max(0, \min(v, 255)).$$

The experiment considers three levels of noise intensity, based on λ, which determine the extent to which the noised segment remains discernible::

- $\lambda = 0.0$ **(Baseline):** No noise is applied.
- $\lambda = 0.5$ **(Medium Noise):** Moderate noise is applied, affecting 50% of the pixel distribution in R_B.
- $\lambda = 1.0$ **(High Noise):** Maximum noise is applied, occluding the entire bounding box and removing all visual information.

Perhaps explain that the idea here is that you have noise levels that allow for various degrees of discernment of what's in the noised segment

5.3 Noise Area Conditions

The study introduces three distinct noise injection area settings:

- **Target Noise:** Noise is applied inside the bounding box, directly affecting the object.
- **Context Noise:** Noise is added outside the bounding box, distorting the surrounding scene while keeping the object intact.
- **All Noise:** Noise is applied both inside and outside the bounding box, affecting the entire image.

These conditions allow us to analyze whether the model relies on the object's visual integrity or contextual information for recognition.

5.4 Semantic Violation Conditions

Semantic violation conditions are defined based on the degree of semantic relatedness between an object and the scene in which it appears. These conditions assess the model's ability to correctly recognize objects when their contextual appropriateness varies. The COOCO dataset provides a structured approach to manipulating semantic coherence, distinguishing objects by their level of relatedness to the scene type. The conditions are defined as follows:

- **High Relatedness Condition or Original Condition**: We assume that since the image is not modified, the object is semantically appropriate for the scene, fitting naturally within the expected context.
- **Medium Relatedness Condition**: The object has a moderate semantic connection to the scene but is not typically expected.
- **Low Relatedness Condition**: The object is highly incongruent with the scene, violating semantic expectations. Recognition in this condition evaluates the model's robustness to semantic anomalies and its reliance on contextual cues.

These conditions allow us to investigate how VLMs handle semantic inconsistencies and whether their object recognition performance degrades as the violation increases.

5.5 Metrics

The model's performance is assessed using several evaluation metrics, with `RefCLIPScore` [20] being a key measure. `RefCLIPScore` quantifies the alignment between model-generated captions, reference labels, and the image by leveraging CLIP-derived similarity. Additionally, we measure simple text-based semantic similarity and accuracy using two variations: a "hard" version based on the Levenshtein similarity ratio and a "soft" version that leverages text-based semantic similarity to evaluate the percentage of correctly identified objects.

5.5.1 RefCLIPScore

`RefCLIPScore` is an evaluation metric that extends `CLIPScore` [20] by incorporating reference captions to assess the quality of generated image captions more comprehensively. It measures both image relevance and linguistic similarity to references by computing the harmonic mean between two components: (1) the `CLIPScore`, which captures the semantic alignment between the generated caption and the image using CLIP [47] embeddings, and (2) the highest cosine similarity between the generated caption and any reference caption. By using the harmonic mean, `RefCLIPScore` ensures that a caption is penalized if it excels in only one aspect while failing in the other, leading to a more balanced assessment. The following details the steps used.

Text and Image Embeddings

Let f_{img} and f_{text} denote the CLIP [47] embedding functions for images and text, respectively. Given an image I and a text T, their corresponding normalized embeddings are:

$$v = \frac{f_{\text{img}}(I)}{\|f_{\text{img}}(I)\|}$$

$$t = \frac{f_{\text{text}}(T)}{\|f_{\text{text}}(T)\|}$$

where v represents the image embedding and t represents the text embedding.

Cosine Similarity

For two normalized embeddings $x, y \in \mathbb{R}^d$, the cosine similarity is computed as:

$$S(x, y) = x^\top y$$

CLIPScore

The `CLIPScore` measures the alignment between a candidate caption c and the image I:

$$\text{CLIP-S}(c, I) = w \cdot \max(S(t_c, v), 0)$$

where $t_c = f_{\text{text}}(c)$ is the embedding of the candidate caption, and w is a weight scaling factor (default $w = 2.5$).

Reference CLIP Score

RefCLIPScore extends CLIP-S by incorporating reference captions. To compute it, we first extract vector representations of all reference captions using CLIP's text encoder, forming the reference embedding set R. In our case, there is a single reference caption, corresponding to the target name from the dataset. The RefCLIPScore is defined as the harmonic mean of the standard CLIP-S score and the highest cosine similarity between the candidate caption, our model's output, and any reference caption:

$$\text{RefCLIP-S}(c, R, v) = \text{H-Mean}(\text{CLIP-S}(c, v), \max(\max_{r \in R} \cos(t_c, t_r), 0)) \tag{5.1}$$

where t_c and t_r are the CLIP text embeddings of the candidate and reference captions, respectively.

In our experiment, we analyze the following three embeddings:

$$v_B = \frac{f_{\text{img}}(R_B)}{\|f_{\text{img}}(R_B)\|}$$

$$t_O = \frac{f_{\text{text}}(O)}{\|f_{\text{text}}(O)\|}$$

$$t_t = \frac{f_{\text{text}}(T_t)}{\|f_{\text{text}}(T_t)\|}$$

where R_B is the image patch corresponding to the bounding box region, O corresponds to the model's predicted output, and T_t is the label of the target. This allows us to evaluate how well the model aligns the image patch representation with the expected textual description. Following [20], which found that prefixing candidates with "A photo depicts" slightly improved correlation, we apply a similar normalization approach to O and T_t. We first strip leading whitespace, convert text to lowercase, and remove definite articles ("a" or "an") if present at the beginning. We then prepend "A photo depicts," selecting "a" or "an" based on the initial letter of the target and output to ensure grammatical correctness.

The final formulation of `RefCLIPScore` is:

$$\text{RefCLIP-S}(t_O, t_t, v_B) = \text{H-Mean}(\text{CLIP-S}(t_O, v_B), \max(\cos(t_O, t_t), 0)) \tag{5.2}$$

5.5.2 Text-Based Semantic Similarity Score

We used the text similarity score to directly measure the semantic closeness between both the output and target label and the output and the scene label, without taking into account the image patch:

$$\text{Text-Sim}(O, T_t) = S(t_O, t_t)$$

and

$$\text{Text-Sim}(O, T_s) = S(t_O, t_s)$$

where T_s its the scene label and t_s is it's text embedding.

5.5.3 Accuracy

Given a dataset of generated labels and reference labels, we define two accuracy metrics: soft accuracy and hard accuracy.

Soft Accuracy

Soft accuracy is based on the cosine similarity between the text embeddings of the generated caption and the reference caption. Let O and T_t be the generated caption and the reference caption, respectively. The soft accuracy is defined as:

$$A_{\text{soft}}(O, T_t) = \begin{cases} 1, & \text{if Text-Sim}(O, T_t) >= 0.9 \\ 0, & \text{otherwise} \end{cases}$$

We chose to base this accuracy metric on text similarity rather than *refCLIPScore*, as we found it to be more interpretable.

Hard Accuracy

Hard accuracy is computed using the Levenshtein similarity ratio, which quantifies the character-level edit distance between the generated caption and the reference caption. The Levenshtein similarity ratio is defined as:

$$\text{Ratio}(s_1, s_2) = 1 - \frac{D(s_1, s_2)}{|s_1| + |s_2|}$$

where s_1 and s_2 are the input sequences, $D(s_1, s_2)$ represents the Levenshtein distance—the minimum number of insertions, deletions, or substitutions required to transform s_1 into s_2—and $|s_1|$, $|s_2|$ denote the lengths of the respective sequences. The similarity ratio ranges from 0 to 1, where $\text{Ratio}(s_1, s_2) = 1$ indicates identical sequences, and $\text{Ratio}(s_1, s_2) = 0$ signifies no shared elements between s_1 and s_2. A value of 0.55 indicates words that are similar, but we have empirically observed that it also includes cases where the target is mentioned along with an additional adjective, which models often tend to add. To exclude instances where the models did not strictly follow the instructions and generated an output that described the target object or scene rather than simply naming it, we excluded from the hard accuracy calculation any output strings exceeding the length of the longest target label (that is 16 characters). This filter is not necessary for the soft accuracy calculation, as it does not focus on string similarity but rather on semantic content. The hard accuracy is then defined as:

$$A_{\text{hard}}(O, T_t) = \begin{cases} 1, & \text{if Ratio}(O, T_t) >= 0.55 \\ 0, & \text{otherwise} \end{cases}$$

where T_c is the generated caption, T_r is the reference caption, and $R_{\text{lev}}(T_c, T_r)$ denotes the Levenshtein similarity ratio between them.

Chapter 6

Results

In this section, we present the experimental results. The first part focuses on `refCLIPScore` and text-similarity outcomes, while the second part examines hard and soft accuracy metrics.

6.1 RefCLIPScore and Text-Similarity

Figure 6.1 illustrates the average `refCLIPScore` for all models under the 0-noise condition, serving as a baseline. Notably, the Qwen2-VL model exhibited the lowest performance, leading to its removal from subsequent analysis. Overall, performance appears to be influenced by the level of relatedness between the target and the scene. Most models show better alignment between the image and the target caption as relatedness increases, except for LLaVA-OneVision and Molmo, which perform worse under medium-relatedness compared to low-relatedness conditions.

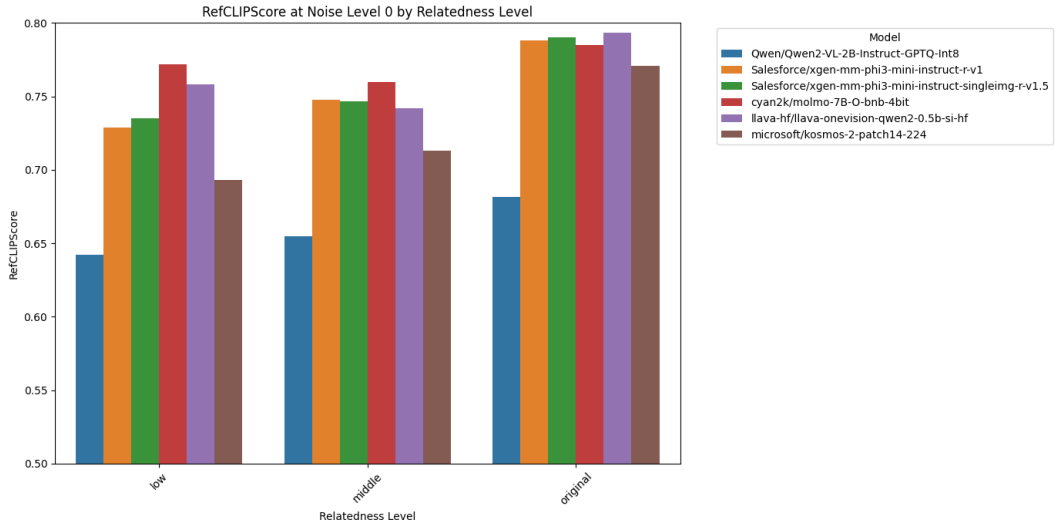


Figure 6.1: Average `refCLIPScore` for all models under the 0-noise condition.

Since our primary objective is to analyze general performance trends across Vision-Language Models (VLMs) rather than identifying the best-performing model, we aggregate

the scores across all evaluated models. Appendix A.1 includes tables presenting the results on the whole dataset for each evaluated model. Figure 6.2 presents the average `refCLIPScore` results, categorized by noise level, relatedness level, and noise area. The same results are present in the first column of Table 6.1. To examine whether image quality influences the results, we conducted the same analysis on the *COOCO* manually filtered subset, with higher-quality generated images. The results of `refCLIPScore` and text-based semantic similarity for both datasets are shown in Table 6.1.

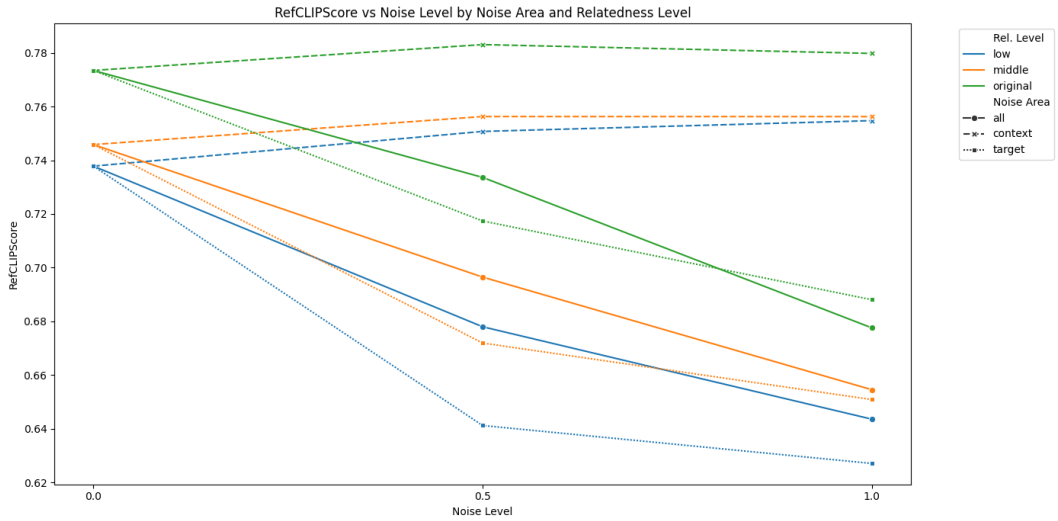


Figure 6.2: Average `refCLIPScore` across different noise conditions, noise areas, and relatedness levels.

6.2 Accuracy

Figure 6.3 presents the hard accuracy results across all experimental conditions. Table 6.2 provides a quantitative summary of both soft and hard accuracy scores, further differentiating performance across unfiltered and manually filtered datasets. The results indicate that as noise levels increase, accuracy generally declines. Notably, the "context" noise area maintains better performance compared to "all" and "target", presenting the highest scores, enforcing the idea of the context as a distractor. Additionally, the "original" relatedness level consistently outperforms the "low" and "middle" levels, suggesting that these models have learned a more robust representation. In the "original" (high-relatedness) condition, maximum noise on the target does not result in a complete drop in accuracy, unlike in other conditions. This may be due to the model's ability to use the context to inform its decisions, where the object is strongly related to the surrounding context. Consequently, the context plays a crucial role in maintaining accuracy despite target noise. Filtering the data (i.e., removing poor-quality images) boosts accuracy across all conditions, underscoring the impact of image quality. Overall, the results show that noise localized on context is less harmful, and high-quality, strongly related training examples yield superior performance.

Rel. Level	Noise Level	Noise Area	refCLIPScore	text sim.	refCLIPScore Subset	text sim. Subset
low	0.0	–	0.738	0.860	0.738	0.860
low	0.5	all	0.678	0.813	0.672	0.808
low	0.5	context	0.751	0.869	0.750	0.868
low	0.5	target	0.641	0.785	0.632	0.777
low	1.0	all	0.644	0.800	0.637	0.794
low	1.0	context	0.755	0.871	0.754	0.871
low	1.0	target	0.627	0.777	0.620	0.769
middle	0.0	–	0.746	0.875	0.743	0.875
middle	0.5	all	0.696	0.838	0.692	0.835
middle	0.5	context	0.756	0.882	0.754	0.883
middle	0.5	target	0.672	0.818	0.671	0.817
middle	1.0	all	0.655	0.820	0.650	0.816
middle	1.0	context	0.756	0.881	0.753	0.882
middle	1.0	target	0.651	0.804	0.651	0.803
original	0.0	–	0.773	0.944	0.776	0.949
original	0.5	all	0.734	0.901	0.731	0.897
original	0.5	context	0.783	0.955	0.786	0.960
original	0.5	target	0.717	0.882	0.723	0.887
original	1.0	all	0.678	0.856	0.673	0.849
original	1.0	context	0.780	0.949	0.781	0.950
original	1.0	target	0.688	0.853	0.698	0.860

Table 6.1: Comparison of `refCLIPScore` and text-based semantic similarity scores across different relatedness levels, noise levels, and noise areas. The last two columns show experiment results on the manually filtered subset, with higher image quality.

6.2.1 Scene-Output Similarity

We categorized all outputs as Correct or Incorrect using the hard accuracy metric, then measured the text-based semantic similarity between the scene label and the output for each group. Table 6.3 reports both scene-output and target-output similarity scores across all experimental conditions, while Figure 6.4 visualizes the scene-output scores by correctness. At the original (high-relatedness) relatedness level, correct outputs generally exhibit higher semantic alignment with the scene than incorrect outputs. However, at middle and low relatedness levels, incorrect outputs often appear more scene-aligned, likely because the target is less closely tied to the scene and the model relies more on scene cues when producing incorrect responses. Notably, applying noise to the target makes both correct and incorrect outputs more scene-oriented, particularly at low relatedness, suggesting that the few correct responses arise because the target retains greater scene relevance than expected, thereby demonstrating the model's reliance on contextual cues. In contrast, when noise is limited to the context, scene similarity decreases as relatedness drops, indicating that outputs become more anchored to the remaining target data.

6.3 Key Findings

Overall, three principal themes emerge from the results:

1. **Influence of Semantic Relatedness:** Across all metrics, images with a target object

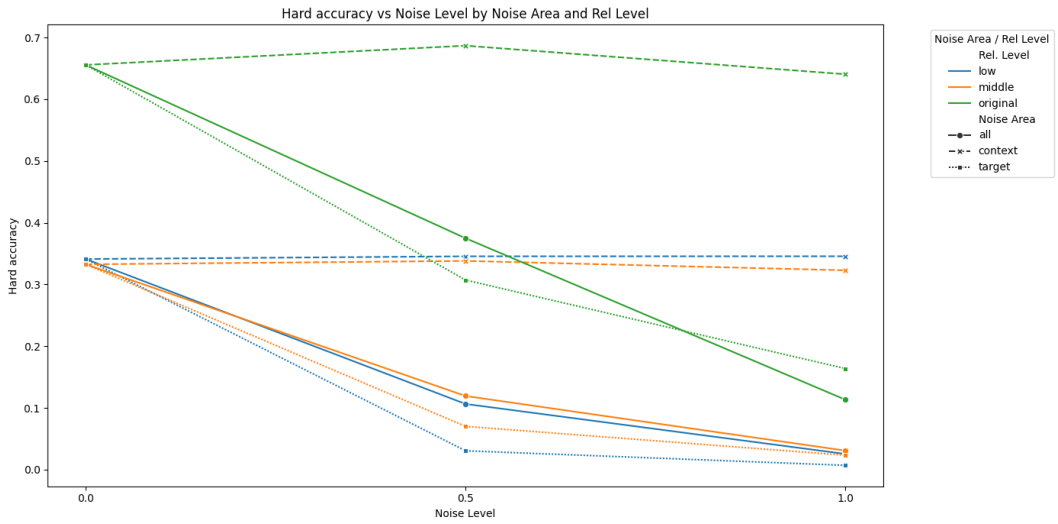


Figure 6.3: Hard accuracy results across all experimental conditions.

highly related to the surrounding scene consistently yielded stronger performance (e.g., higher accuracy and `refCLIPScore`) than those with middle or low related targets. When target and scene were congruent, the models more reliably identified and labeled the target, demonstrating that contextual fit supports referent recognition. Conversely, as semantic relatedness declined, incorrect outputs stayed similar to the broader scene, indicating that the models tended to rely on scene cues when the target's own features offered limited support, degradating the performance.

2. **Role of Noise and Context:** Introducing noise generally reduced accuracy and text-image alignment, especially when the target region was obscured. However, when noise was applied exclusively to the context area, the models performed better, suggesting that partially removing surrounding information may reduce distraction and focus the model on the target. In contrast, when noise was evenly distributed or concentrated solely on the target, the models shifted toward scene-driven outputs. This reliance on context was particularly evident in low-relatedness conditions, where the models generated more scene-oriented descriptions.

3. **Context as Both Facilitator and Distractor:** In high-relatedness settings, contextual cues clearly facilitated identification by compensating for target noise. Yet in scenarios with low relatedness and noise, the same context led to misidentifications. Outputs became broadly scene-aligned but lost fidelity in identifying the actual target. These dual effects highlight that context can be a resource or a confounding influence, depending on how strongly the target ties in with the scene.

We will further discuss these findings in the Analysis Chapter 8.

Rel. Level	Noise Level	Noise Area	Soft Acc.	Hard Acc.	Soft Acc. Filt.	Hard Acc. Filt.
low	0.0	–	0.339	0.341	0.362	0.383
low	0.5	all	0.098	0.106	0.098	0.122
low	0.5	context	0.356	0.346	0.373	0.382
low	0.5	target	0.029	0.031	0.027	0.031
low	1.0	all	0.024	0.025	0.022	0.027
low	1.0	context	0.362	0.346	0.380	0.384
low	1.0	target	0.004	0.007	0.002	0.005
middle	0.0	–	0.351	0.333	0.361	0.355
middle	0.5	all	0.136	0.119	0.134	0.134
middle	0.5	context	0.366	0.338	0.373	0.365
middle	0.5	target	0.076	0.070	0.079	0.083
middle	1.0	all	0.040	0.031	0.035	0.035
middle	1.0	context	0.354	0.323	0.364	0.350
middle	1.0	target	0.023	0.023	0.025	0.030
original	0.0	–	0.749	0.656	0.752	0.677
original	0.5	all	0.447	0.375	0.398	0.359
original	0.5	context	0.799	0.687	0.803	0.711
original	0.5	target	0.361	0.307	0.382	0.340
original	1.0	all	0.151	0.113	0.112	0.096
original	1.0	context	0.758	0.640	0.743	0.647
original	1.0	target	0.199	0.164	0.245	0.210

Table 6.2: Results Table with hard and soft accuracy scores for both whole and manually filtered datasets, across all experimental conditions.

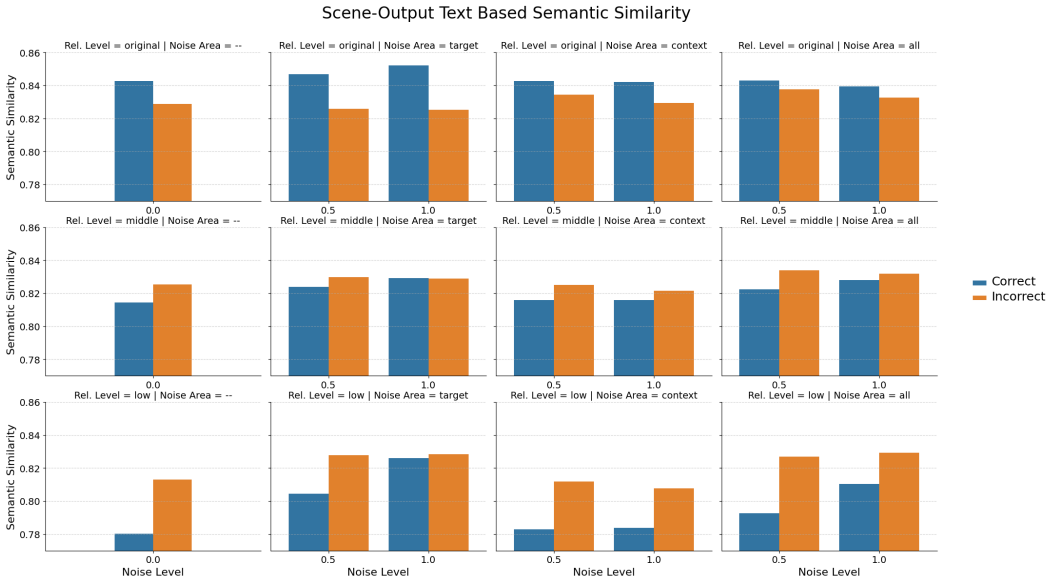


Figure 6.4: Scene-output text-based semantic similarity values across all experimental conditions. The comparison highlights differences between correct and incorrect outputs.

Noise Area	Noise Level	Rel. Level	Target/Out.		Scene/Out.	
			Correct	Incorrect	Correct	Incorrect
–	0.0	low	0.967	0.812	0.780	0.813
		middle	0.970	0.836	0.815	0.825
		original	0.990	0.876	0.843	0.829
all	0.5	low	0.946	0.800	0.793	0.827
		middle	0.953	0.824	0.822	0.834
		original	0.990	0.853	0.843	0.838
	1.0	low	0.910	0.799	0.810	0.829
		middle	0.922	0.818	0.828	0.832
		original	0.977	0.842	0.839	0.832
context	0.5	low	0.968	0.821	0.783	0.812
		middle	0.969	0.842	0.816	0.825
		original	0.990	0.887	0.843	0.834
	1.0	low	0.970	0.824	0.784	0.808
		middle	0.969	0.842	0.816	0.821
		original	0.990	0.880	0.842	0.829
target	0.5	low	0.937	0.787	0.805	0.828
		middle	0.946	0.816	0.824	0.830
		original	0.984	0.849	0.847	0.826
	1.0	low	0.846	0.784	0.826	0.828
		middle	0.900	0.809	0.829	0.829
		original	0.965	0.843	0.852	0.825

Table 6.3: Scene-output and target-output text-based semantic similarity values across all experimental conditions. Correct and incorrect output categories are evaluated separately.

Chapter 7

Attention Deployment Analysis

In this chapter, we explore how vision-language models manage their attention mechanisms during scene processing. Our focus is on LLaVA-OneVision [29], which stood out as one of the top performers in our experiments. Additionally, there are publicly available resources for this model family, such as LLaVA-CAM [75] for LLaVA 1.5 [32], that streamline the visualization of attention layers. Notably, the open-source toolkit available at `https://github.com/zjysteven/VLM-Visualizer` offers direct implementation (with minimal adjustments, given it was designed for prior versions of the LLaVA family), making it an ideal choice for a deeper exploration of how visual cues are integrated into the referential language generation process.

7.1 LLaVA-OneVision Architecture and Input Processing

LLaVA-OneVision [29] builds upon the LLaVA framework [34] by incorporating a vision-language alignment mechanism that efficiently integrates image features into a language model. The architecture consists of three main components: a Vision Encoder, a Multimodal Projector, and a Large Language Model (LLM). The Vision Encoder extracts feature representations from input images, the Multimodal Projector maps these features into the word embedding space, and the LLM processes both text and projected image embeddings to generate outputs. Specifically, LLaVA-OneVision adopts SigLIP [72] as its Vision Encoder, denoted as $g_\psi(X_v)$, which transforms an input image X_v into a grid of visual feature representations Z_v. These features are then passed through a two-layer MLP projector $p_\theta(Z_v)$ to produce a sequence of visual tokens H_v, which are compatible with the LLM's word embedding space. For the language backbone, LLaVA-OneVision utilizes Qwen-2 [68] as its LLM $f_\phi(\cdot)$, selected for its strong language modeling and instruction-following capabilities.

7.1.1 Visual Representations and AnyRes Encoding

The representation of visual signals plays a crucial role in the performance of the visual encoding process. Two key factors influence this representation: the resolution in the raw pixel space and the number of tokens in the feature space. These factors define the visual input representation configuration as (resolution, #tokens). Scaling both factors improves performance, particularly in tasks requiring fine-grained visual details. However, to balance

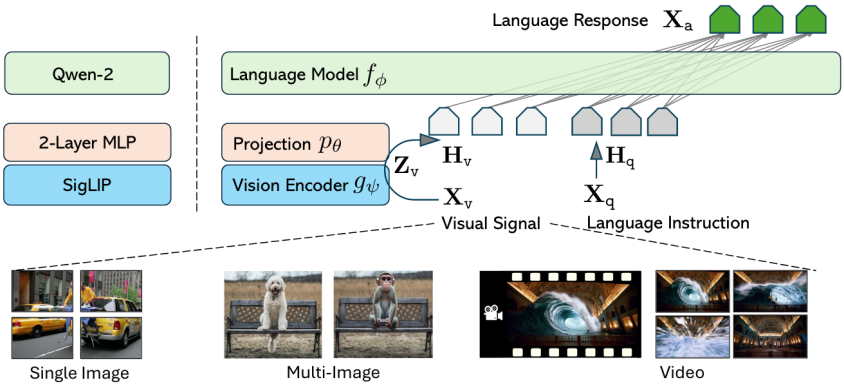


Figure 7.1: LLaVA-OneVision network architecture. Left: The current model instantiation; Right: the general form of LLaVA architecture in [33], but is extended to support more visual signals. Source: [29]

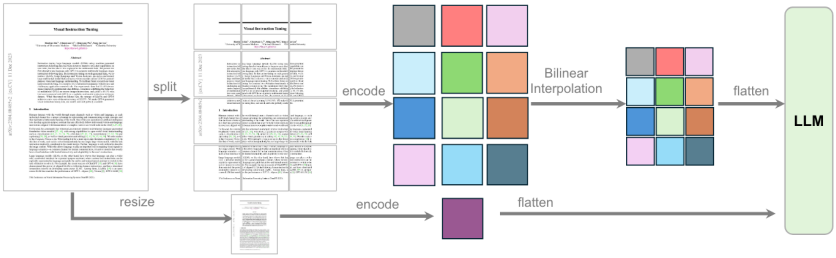
performance and computational cost, empirical evidence suggests that scaling resolution is more effective than increasing the number of tokens. To this end, LLaVA-OneVision adopts an AnyRes strategy with pooling.

For AnyRes with a configuration of width a, height b, it divides the image into $a \times b$ crops, each with the shape (a, b). Each crop has the same resolution suitable for the vision encoder. Assuming there are T tokens per crop, the total number of visual tokens is $L = (a \times b + 1) \times T$, where the base image is resized before being fed into the vision encoder. It considers a threshold τ, and reduces the number of tokens per crop, using bilinear interpolation if needed:

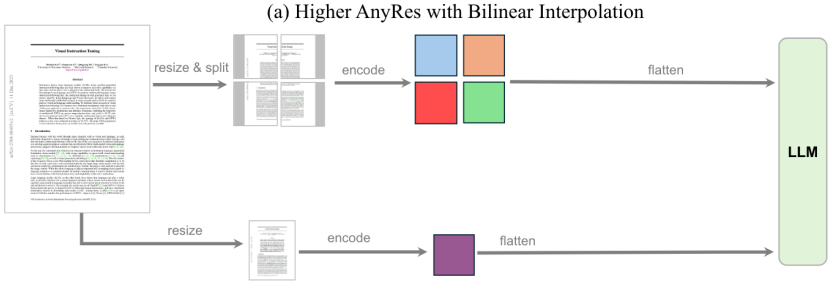
$$T_{\text{new}} = \begin{cases} \frac{\tau}{(a \times b + 1)}, & \text{if } L > \tau \\ T, & \text{if } L \leq \tau \end{cases} \tag{7.1}$$

A set of spatial configurations (a, b) is defined to specify various methods for cropping images, thereby accommodating images of different resolutions and aspect ratios. Among them, the configuration that requires a minimum number of crops is selected. Specifically, they employ the AnyResMax-9 [29] strategy. Using SO400M [72] as the Vision Encoder, each input image (or grid) is processed into 729 visual tokens. Consequently, the maximum number of visual tokens for a single image is $729 \times (1 + 9)$, where 1×729 represents the base tokens and 9×729 accounts for the grid tokens. Importantly, if bilinear interpolation is used to reduce the number of tokens in the image crops, the base image tokens remain constant at 729. AnyRes encoding allows flexible visual token allocation by adjusting the spatial resolution and token distribution across different visual tasks. The AnyRes strategy follows these principles:

- **Single-Image Representation**: A large maximum spatial configuration (a, b) is chosen to maintain the original image resolution without resizing. A higher number of visual tokens per image ensures a long sequence that captures detailed visual information, supporting transfer from image-based training to video tasks.
- **Multi-Image Representation**: For multiple images, only the base resolution is maintained. The vision encoder extracts feature maps directly without performing multi-cropping, optimizing computational efficiency.



(a) Higher AnyRes with Bilinear Interpolation



(b) The original AnyRes

Figure 7.2: The visual representations. Top: The new Higher AnyRes scheme with Bilinear Interpolation to deal with images of higher resolution; Bottom: the original AnyRes. Source: [29]

- **Video Representation**: Each video frame is resized to the base image resolution before being processed by the vision encoder. To manage the computational cost, bilinear interpolation is applied to reduce the number of tokens per frame, enabling the model to accommodate a greater number of frames while maintaining a balance between performance and cost.

This design strategy aims to balance the data from various modalities, ensuring a more equitable representation that is transferable from the perspective of the language model. For instance, a high-resolution image can be interpreted as a composition of multiple images, and multiple images can be understood as a shorter video. The visualization of the token allocation strategy just described is given in Figure 7.3

7.1.2 Input Processing and Concatenation Strategy

The input to LLaVA-OneVision consists of both text tokens and vision features, which are combined using a structured concatenation strategy. The process follows these steps:

1. **Image Encoding**: The model first encodes the input image X_v using the SigLIP vision encoder $g_\psi(X_v)$. The output is a grid of image features Z_v, capturing spatial and semantic information.

2. **Projection to Token Space**: The image features Z_v are projected into the LLM's word embedding space using a two-layer MLP $p_\theta(Z_v)$, resulting in a set of visual tokens H_v. These tokens serve as a compact representation of the visual content in the same dimensional space as text embeddings.

Figure 7.3: The visual representation strategy to allocate tokens for each scenario in LLaVA-OneVision. The maximum number of visual tokens across different scenarios is designed to be similar, ensuring balanced visual representations to accommodate cross-scenario capability transfer. Note that 729 is the #tokens for SigLIP to encode a visual input of resolution 384×384. Source: [29]

3. **Spatial Pooling and Token Reduction**: Given the AnyRes strategy, if the number of image tokens exceeds the predefined threshold τ, a pooling mechanism is applied to reduce spatial redundancy. For images with a high token count, bilinear interpolation downscales the representation while maintaining key spatial structures. After this step, the final token set H'_v is obtained, ensuring that the visual input remains computationally efficient while retaining meaningful semantic details.

4. **Text Tokenization and Embedding**: The input text X_q is tokenized into discrete indices input_ids and embedded using the LLM's embedding layer, yielding text embeddings $H_q = f_\phi(X_q)$.

5. **Insertion of Image Tokens**: The text sequence contains special token placeholders, indicating where visual information should be injected. The model iterates through the tokenized input, replacing these placeholders with the corresponding visual tokens H'_v.

6. **Masking and Padding**: To ensure correct training behaviour, labels for image tokens are masked, preventing loss computation on these tokens. The sequence is then padded to maintain a consistent length.

The final input sequence to the transformer consists of interleaved text embeddings H_q and image embeddings H_v. The model processes this combined sequence as a single input stream, enabling multimodal reasoning. During inference, the LLM generates textual outputs based on the fused representations, ensuring that visual information remains contextually grounded throughout the response generation process.

Figure **??** shows an example of the standard input text template used with this model. The System's text, which remains consistent across all prompts, is highlighted in `red`, while the user's input is marked in `green`. Image tokens replace the special placeholder `<image>`, which is highlighted in `blue`.

A visual depiction of the final interleaved input matrix is presented in Figure 7.5, illustrating how text and vision embeddings are structured for processing.

Input Text Template Example

```
<|im_start|> system
You are a helpful assistant.<|im_end|>
<|im_start|> user
<image>
[User Prompt]
<|im_end|>
<|im_start|> assistant
```

Figure 7.4: Example of the conversational text template used for the LLaVa-OneVision model.

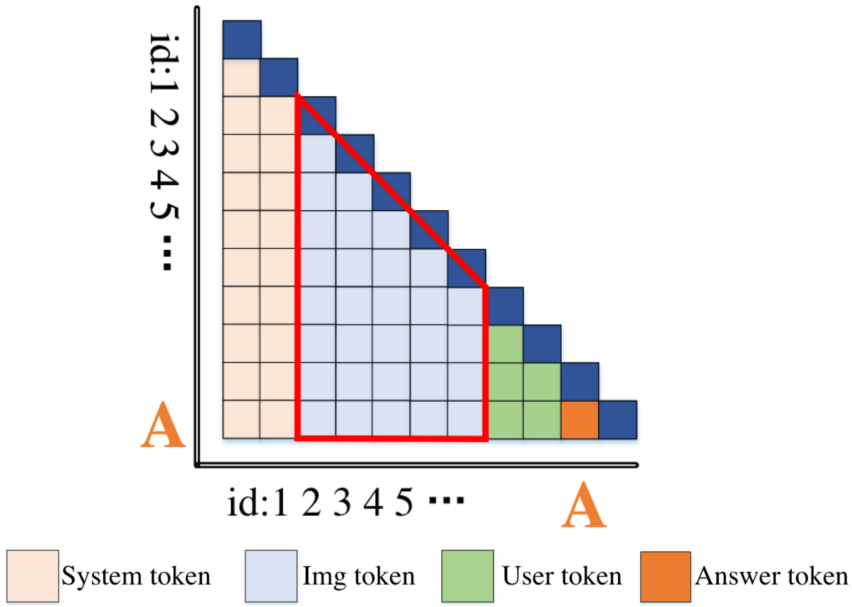


Figure 7.5: This image illustrates the composition of an interleaved matrix in multimodal models. System tokens are shown in red, Image tokens in blue, User tokens in green, and Answer tokens, appended after each iteration of the generative decoder model, in orange. Source: [75]

7.2 Attention Aggregation

In multi-head attention architectures, each attention head learns specialized focus patterns, often resulting in a complex distribution of weights. By aggregating these weights across layers and heads, we obtain a more interpretable representation that highlights overall attention trends. This section presents our approach to attention aggregation in both language models and vision transformers. These aggregated views allow us to capture key interactions in the model's focus, supporting more intuitive interpretations of its decision-making process.

7.2.1 Self-Attention

In the transformer model [58] given an input x, the self-attention mechanism assigns to each token x_i a set of attention weights over the tokens in the input:

$$\text{Attn}(x_i) = (\alpha_{i,1}(x), \alpha_{i,2}(x), ..., \alpha_{i,i}(x)) \tag{7.2}$$

where $\alpha_{i,j}(x)$ is the attention that x_i pays to x_j. The weights are positive and sum to one. In the multi-layer, multi-head setting, α is specific to a layer and head. The attention weights $\alpha_{i,j}(x)$ are computed from the scaled dot-product of the query vector of x_i and the key vector of x_j, followed by a softmax operation. The attention weights are then used to produce a weighted sum of value vectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7.3}$$

using query matrix Q, key matrix K, and value matrix V, where d_k is the dimension of K. In a multi-head setting, the queries, keys, and values are linearly projected h times, and the attention operation is performed in parallel for each representation, with the results concatenated.

When analyzing attention patterns, we extract attention matrices that capture the distribution of focus assigned by the model. These matrices contain the softmax-normalized attention weights:

$$\text{Attention Weights}(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{7.4}$$

7.2.2 Attention Aggregation

The attention analysis we perform involves aggregating attention across all heads within each layer to produce a comprehensive view of how tokens interact [60]. Given that our framework follows a vision-language architecture, this process is conducted separately for both the Vision Encoder and the Large Language Model (LLM) Decoder attention matrices.

Aggregating LLM Attention

To obtain an interpretable view of the attention across layers and heads, we aggregate attention values in a structured manner. The first step in this process is computing the mean attention weight across all heads for each layer. Given an attention tensor A of shape (L, H, N, N), where L represents the number of layers, H the number of heads, and N the number of tokens, we compute the mean attention per layer as:

$$\bar{A}_l = \frac{1}{H}\sum_{h=1}^{H} A_{l,h} \tag{7.5}$$

To enhance interpretability, we follow a common practice of nullifying the attention to the first token (e.g., beginning-of-sequence, BOS), as this token typically absorbs disproportionate attention due to its special role in autoregressive models [60]. Specifically, we set:

$$\bar{A}_{l,i,0} = 0 \quad \forall i > 0 \tag{7.6}$$

ensuring that the BOS token does not interfere with the analysis. The normalized attention weights are then computed as:

$$\hat{A}_l^{(i,j)} = \frac{\bar{A}l^{(i,j)}}{\sum_{j=1}^{N} \bar{A}_l^{(i,j)}} \tag{7.7}$$

where N is the sequence length. The final aggregated attention matrix is obtained by averaging across layers:

$$A_{\text{LLM}} = \frac{1}{L} \sum_{l=1}^{L} \hat{A}_l \tag{7.8}$$

where L is the total number of layers. Our method explicitly distinguishes between prompt tokens and generated tokens when aggregating attention. This distinction is fundamental because prompt tokens receive attention from all subsequent tokens, while generated tokens adhere to an autoregressive dependency structure, attending only to past tokens. To compute the final attention representation, we first process the prompt tokens' attention matrix by aggregating it across heads, normalizing it, and averaging it across layers. A similar process is then applied iteratively to the generated tokens, with the following modifications: only the last row, corresponding to the newly generated token, is retained, and attention to the beginning-of-sequence (BOS) token is nullified. Once computed, each resulting vector is stacked with the aggregated prompt token matrix to construct the final representation. Given that the prompt and generated sequences can have different lengths, we pad the generated token attention vectors to a common length before stacking, ensuring alignment:

$$A_{\text{final}} = \text{stack}(A_{\text{prompt}}, \text{pad}(A_{\text{output}})) \tag{7.9}$$

Aggregating Attention in Vision Transformers (ViTs)

In vision transformer models, attention mechanisms operate over spatial token representations, typically corresponding to image patches. To analyze attention patterns effectively, we compute a layer-wise average attention matrix across the dataset as a baseline for evaluating individual inputs. We subtract this baseline to remove systematic attention biases, ensuring that the remaining attention reflects image-specific patterns rather than inherent model tendencies. To prevent confounding effects related to image size, we excluded all images with dimensions other than 640x640 from the attention deployment analysis (that were a small subset of the dataset).

The average attention matrix is computed in two steps:

1. **Per-Trial Layer Aggregation**: For each trial, we first select a specific layer and compute the mean attention across all heads:

$$\bar{A}_l^{(t)} = \frac{1}{H} \sum_{h=1}^{H} A_{l,h}^{(t)} \tag{7.10}$$

where $A_{l,h}^{(t)}$ represents the attention matrix for head h in layer l for trial t. To normalize attention values within the selected layer, we divide by the sum of attention weights across all tokens:

$$A_{\text{norm},l}^{(t)} = \frac{\bar{A}_l^{(t)}}{\sum_j \bar{A}_l^{(t)}[j]} \tag{7.11}$$

This normalization ensures that attention values are comparable across different trials.

2. **Dataset-Wide Averaging Across Trials**: Once the normalized attention maps for each trial have been computed, they are aggregated across all trials to obtain the dataset-wide average attention for each layer:

$$A_{\text{avg},l} = \frac{1}{T} \sum_{t=1}^{T} A_{\text{norm},l}^{(t)} \tag{7.12}$$

where T is the total number of trials. This process ensures that the average attention baseline is computed separately for each layer.

The final average attention matrices, one for each layer, provide a reference baseline for interpreting model attention behaviour.

The following aggregation process begins with the same two steps described in per-trial layer aggregation. The adjusted attention map is then computed by subtracting the precomputed average attention matrix for that layer and applying a rectified linear activation (ReLU):

$$A_{\text{adjusted},l}^{(t)} = \text{ReLU}(A_{\text{norm},l}^{(t)} - A_{\text{avg},l}) \tag{7.13}$$

This operation preserves only positive deviations from the dataset-wide mean, emphasizing attention shifts that are particularly relevant for the given trial. Thus, the final aggregated attention representation for the selected layer is:

$$A_{\text{ViT},l}^{(t)} = A_{\text{adjusted},l}^{(t)} \tag{7.14}$$

By focusing on deviations from the dataset-wide baseline, this approach provides an interpretable view of how vision transformers dynamically adjust their attention to different spatial regions, rather than merely capturing absolute attention magnitudes.

7.3 Attention Over Image Regions

In this section, we describe the computation of attention over image regions, which helps us analyze how the model distributes its focus when processing visual inputs. The attention scores over different areas of the image can provide insights into the model's interpretability and decision-making process.

We define the grid size of the visual input as g, obtained from the vision tower of the model, which, in the case of SO400M [72], is equal to 27. The image is divided into $g \times g$ patches, each corresponding to a vision token. The attention computation involves the set of output tokens O, excluding the final token (`<|im_end|>` token):

$$O = \{o_1, o_2, ..., o_{|O|-1}\}. \tag{7.15}$$

For each output token o_i, the attention scores over vision tokens are extracted from the attention matrix of the language model, denoted as $A_{i,j}$, where j indexes the vision tokens and i the matrix row corresponding to the generated output token o_i. We consider only the base image tokens, which remain constant at 729, and exclude the ones related to image crops. This ensures that the analysis focuses on the attention over the entire image, leveraging a more holistic representation. These scores are then normalized to sum to 1:

$$\hat{A}_{i,j} = \frac{A_{i,j}}{\sum_{j \in J} A_{i,j}}, \tag{7.16}$$

where J represents the set of vision tokens. Each vision token has a corresponding attention map, denoted by V_j, which is reshaped to the grid dimensions (g, g). The weighted sum of these maps forms the attention over the image for a given output token:

$$A_i^{img} = \sum_{j \in J} \hat{A}_{i,j} \cdot V_j. \tag{7.17}$$

To compute the final attention map over the image for the whole output, we sum the individual token-level maps and divide them by the number of considered tokens:

$$A_{final}^{img} = \frac{1}{|O| - 1} \sum_{i \in O} A_i^{img}. \tag{7.18}$$

This aggregated map is then upsampled to match the original image dimensions using nearest-neighbor interpolation.

7.3.1 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 region. The total attention over the target area is computed as:

$$a_{target} = \sum_{(x,y) \in B} A_{final}^{img}(x, y). \tag{7.19}$$

Similarly, the attention over the context region, excluding both the target and irrelevant padding areas, is given by:

$$a_{context} = \sum_{(x,y) \in C} A_{final}^{img}(x, y), \tag{7.20}$$

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 $A_{\mathrm{ViT},l}$, ensuring that we obtain an attention distribution over the image, and thus over both the target and context, specific to each layer. Figure A.1 illustrates the attention deployment across the input image, organized by layers and output tokens.

7.3.2 Attention Allocation Ratio

To analyze attention allocation across different levels l of the vision encoder and experimental trials t, we compute the ratio r between the total attention weights assigned to the target (a_{target}) and the context ($a_{context}$):

$$r_l^{(t)} = \frac{a_{target,l}^{(t)}}{a_{context,l}^{(t)}} \tag{7.21}$$

A value of r closer to 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 analyzing its

variations across different experimental conditions. We then compute the mean ratio value for each encoder layer by averaging across all trials:

$$\bar{r}_l = \frac{1}{T} \sum_{t=1}^{T} r_l^{(t)} \tag{7.22}$$

where T denotes the total number of trials. This allows us to assess how attention is distributed across different hierarchical levels of the encoder under varying experimental conditions.

7.4 Results

We analyze attention allocation by segmenting the data according to the experimental conditions: the relatedness level (*original*, *low*, and *middle*), the image region where noise is applied (*target*, *context*, and *all*), and the noise level (0.0, 0.5, and 1.0).

7.4.1 Whole Vs. Manually Filtered Dataset

We first conduct this analysis on both the entire dataset and a manually filtered subset to ensure that variations in image quality do not influence the results. Figure 7.6 presents a visualization of the findings. For the 0.0 noise condition, only one noise application setting is considered, as the absence of noise eliminates any meaningful distinction between regions. Table 7.1 presents the r values averaged across layers. In the 1.0 noise level condition, the condition where noise is applied in the whole image (*all*) is missing because data collection was incomplete due to computational constraints, specifically the exhaustion of available GPU time.

Noise Area	Relatedness Level	Whole Dataset			Subset		
		Noise 0.0	Noise 0.5	Noise 1.0	Noise 0.0	Noise 0.5	Noise 1.0
target	original	0.072	0.067	0.066	0.078	0.076	0.077
context	original	0.072	0.108	0.137	0.078	0.119	0.152
all	original	0.072	0.068	Not Available	0.078	0.081	Not Available
target	middle	0.087	0.081	0.081	0.089	0.090	0.092
context	middle	0.087	0.135	0.160	0.089	0.142	0.167
all	middle	0.087	0.087	Not Available	0.089	0.096	Not Available
target	low	0.091	0.083	0.081	0.097	0.093	0.092
context	low	0.091	0.141	0.166	0.097	0.151	0.178
all	low	0.091	0.092	Not Available	0.097	0.103	Not Available

Table 7.1: Table showing the r values for the whole dataset and the manually filtered subset, averaged across layers for different conditions, relatedness levels, and noise levels.

7.4.2 Correct Vs. Incorrect

We conducted a second analysis, categorizing the output on the whole dataset based on Correct and Incorrect responses according to the hard accuracy metric explained in Section 5.5.3. The mean attention ratio per layer for correct and incorrect responses across the different conditions is shown in Figure 7.7. Additionally, Table 7.2 provides the averaged ratio across layers for all the different conditions.

Noise Area	Relatedness Level	Correct Responses			Incorrect Responses		
		Noise 0.0	Noise 0.5	Noise 1.0	Noise 0.0	Noise 0.5	Noise 1.0
target	original	0.078	0.078	0.060	0.061	0.063	0.068
context	original	0.078	0.114	0.147	0.061	0.094	0.114
all	original	0.078	0.073	Not Available	0.061	0.066	Not Available
target	middle	0.099	0.120	0.093	0.081	0.078	0.081
context	middle	0.099	0.159	0.192	0.081	0.122	0.144
all	middle	0.099	0.113	Not Available	0.081	0.083	Not Available
target	low	0.096	0.114	0.099	0.089	0.083	0.081
context	low	0.096	0.150	0.177	0.089	0.137	0.160
all	low	0.096	0.116	Not Available	0.089	0.089	Not Available

Table 7.2: Averaged ratio across layers for correct and incorrect responses, across different conditions.

7.5 Key Findings

Our analysis of attention allocation across different experimental conditions reveals several trends regarding how the quality of the target image, relatedness levels and noise influence attention patterns.

1. **Image Quality:** Across all conditions, the target-region attention ratio (r) is relatively small but is consistently higher in the manually filtered subset. This indicates that filtering out low-quality or ambiguous images makes the target more salient.

2. **Effect of Relatedness Level:** Regardless of noise configuration, the *low*-relatedness condition yields the highest target-attention ratios, followed by *middle*, with the *original* (high-relatedness) condition having the lowest ratios. This pattern indicates that when objects are semantically incongruent with the scene, the model is forced to rely more on the target object itself, whereas high semantic coherence encourages more contextual integration.

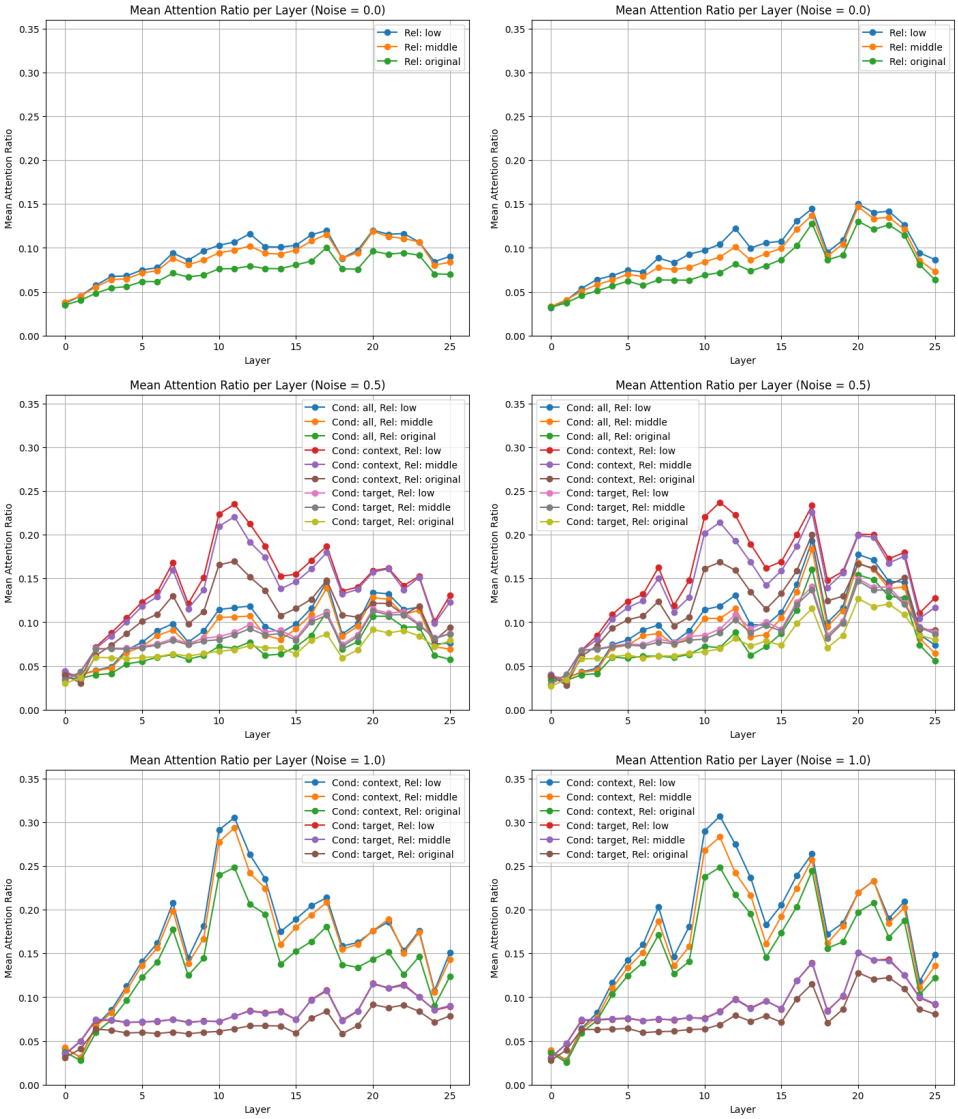
3. **Effect of Noise on Attention Allocation.** Introducing noise to different image regions yields distinct shifts in attention. When noise is placed in the *context*, r exhibits the largest increase, especially at higher noise levels, implying that the model compensates by focusing on the target. By contrast, noise on the *target* generally reduces r, reflecting higher uncertainty in the target region. In general the increase of noise makes these effects more evident. When noise is applied to the *entire* image (at noise level 0.5), there are not clear differences with respect the baseline condition.

4. **Correct Responses:** Correct responses generally exhibit higher attention to the target region, indicating that a successful strategy involves focusing on the target, even when it is affected by noise. The only exception occurs in the *original* condition at high noise levels, where the model relies more on any remaining clean contextual information. This residual context proves useful, as it is highly related to the target.

5. **Target Noise Condition, Correct vs. Incorrect Responses:** For correct responses, attention ratios peak at intermediate noise levels (0.5) before declining. For instance, in the *original* condition, the values are 0.078 at noise 0.0, 0.078 at noise 0.5, and 0.060 at noise 1.0, indicating a flexible allocation of attention to more informative areas. In contrast, incorrect responses exhibit a different pattern. In the *original*

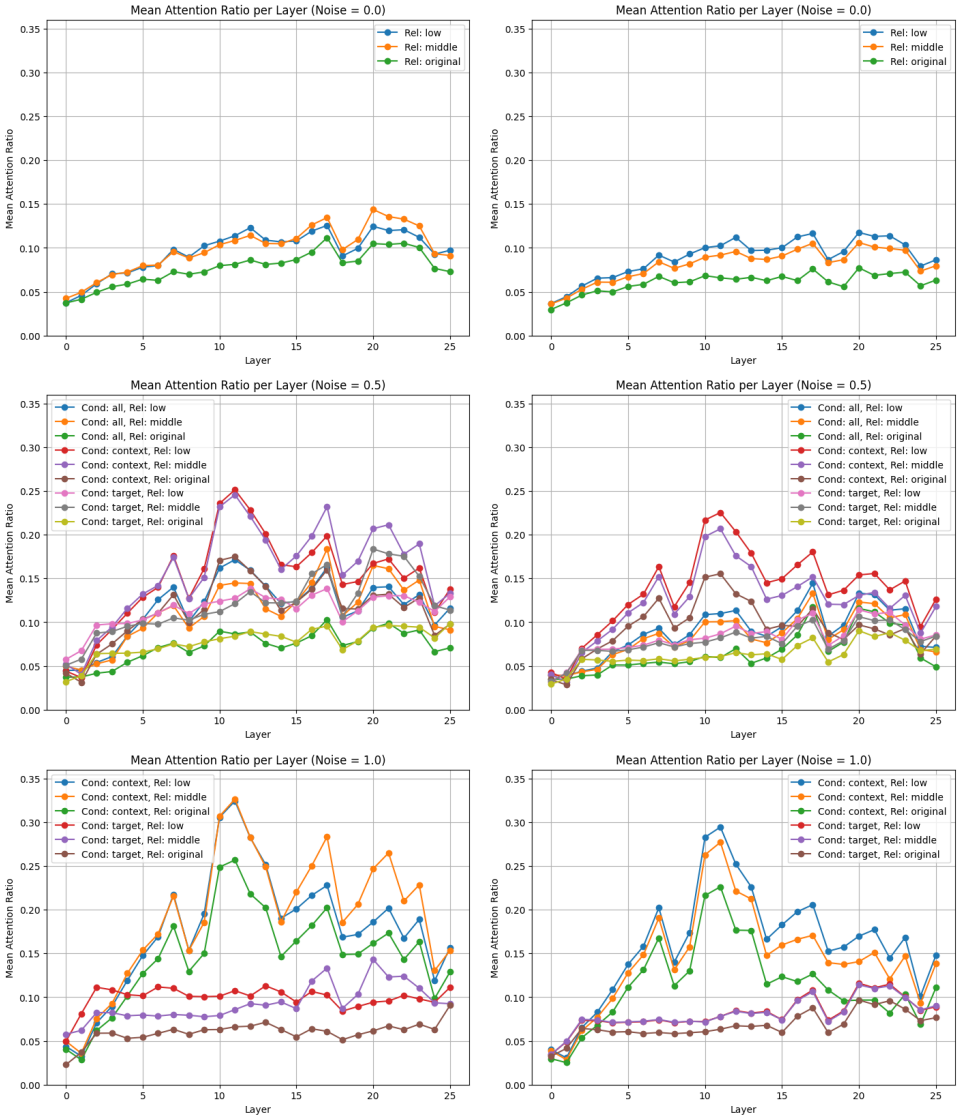
condition, attention ratios gradually increase ($0.061 \rightarrow 0.063 \rightarrow 0.068$), suggesting a failure to redirect attention away from the occluded target. In the *middle* condition, the ratios remain nearly constant, whereas in the *low* condition, they show a slight decrease, which negatively impacts performance since the target is the primary source of useful information in the low-relatedness condition.

We will discuss these findings in the Analysis Chapter 8.



(a) Whole Dataset (b) Manually Filtered Subset

Figure 7.6: Visualization of attention allocation analysis across experimental conditions. The right column displays plots corresponding to the manually filtered subset.



(a) Correct Responses (b) Incorrect Responses

Figure 7.7: Mean ratio per layer for correct and incorrect responses across different conditions.

Chapter 8

Analysis

Below is a discussion of how the key findings from the **Results** and **Attention Analysis** chapters support, or nuance, the four hypotheses regarding context-driven perception and referring expression generation.

8.1 Performance Drop in Low-Relatedness Occlusions

The hypothesis that performance drops significantly when occluded targets exhibit low semantic relatedness is well-supported by the findings. The data shows that high semantic congruence enhances recognition, whereas low relatedness forces reliance on less informative scene cues, leading to degraded performance. Moreover, noise, especially when applied to the target, exacerbates these effects by reducing focused attention, further impairing correct identification. Thus, the evidence indicates that contextual cues are insufficient to compensate for missing visual information in low-relatedness occlusions.

8.2 Scene-Driven Outputs Under Heavy Occlusion

The data provide nuanced support for the hypothesis that under heavy occlusion, outputs should be driven more by the broader scene context rather than by an occluded object. In scenarios where noise is introduced, especially when it obscures the target, the models exhibit a marked shift towards scene-driven outputs, reinforcing the idea that the context becomes the dominant source of information when the object itself is compromised, even if not so informative for the target identity. Additionally, although attention allocation data indicate that correct responses generally involve higher focus on the target, the exception at high noise levels, where a successful strategy involves redirecting attention away from the occluded target towards the scene, especially in the high relatedness condition, further underscores the adaptive benefit of leveraging contextual cues under heavy occlusion. Collectively, these findings suggest that both the semantic fit between object and scene and the strategic redistribution of attention in noisy, occluded environments are critical determinants of model performance, ultimately supporting the view that scene context plays a pivotal role when objects are heavily occluded and semantically misaligned.

8.3 Graded Relatedness Effects

The data strongly support the hypothesis that graded relatedness significantly impacts model performance. When the target object is highly congruent with the scene, models demonstrate higher accuracy and better referent recognition, indicating that a strong contextual fit facilitates identification. Conversely, as semantic relatedness decreases, performance degrades, with models increasingly relying on broader scene cues that can lead to misidentifications. Moreover, attention allocation patterns reveal that low-relatedness conditions prompt a compensatory increase in focus on the target, yet this shift is not sufficient to overcome the detrimental effects of weak semantic ties. Additionally, while noise generally hampers performance, its impact varies depending on whether it disrupts the target or the context, further emphasizing that a coherent scene context is essential for effective target recognition. Collectively, these findings confirm that as object-scene congruence diminishes, so does the model's ability to accurately identify and describe the target.

8.4 Attention Shifts

The hypothesis posits that when the target is heavily distorted or noisy, vision-language models (VLMs) should shift their focus from the target region to contextual cues, much like humans relying on scene-based expectations. The data provide a nuanced perspective: while filtering low-quality images increases the target's saliency (suggesting that a clear target naturally attracts attention), high semantic coherence leads to lower target attention ratios, implying a reliance on context when the target is less informative. Moreover, introducing noise directly to the target decreases the attention ratio, reinforcing the idea that VLMs compensate for degraded target quality by attending more to the surrounding scene. Interestingly, although correct responses generally show higher target-region attention—hinting at the benefits of focusing on the target—this trend reverses in extreme noise conditions, where a flexible reduction in target focus correlates with improved performance. Overall, these findings suggest that while maintaining target attention is often beneficial, an adaptive shift toward contextual elements under heavy noise can enhance task performance, thus partially supporting the hypothesis.

8.5 Final Discussion

Previous studies on VLMs target-context attention deployment [51] and context as a resilience source [24] offer insights that can be reinterpreted in light of the new findings. Schuz et al. [51] demonstrated that a Transformer-based REG model tends to allocate disproportionate attention to the target object, particularly for head nouns, while context information is integrated more variably depending on the linguistic role. The new results extend this by showing that when the target's semantic relatedness to the context is high, the model benefits from a balanced attention distribution that integrates contextual cues, whereas in low-relatedness scenarios, the model is forced to compensate by overly focusing on the target itself. Additionally, the current work's nuanced findings regarding noise, wherein noise in the context can actually help by reducing distraction, but noise on the target undermines performance, resonates with the differential weighting of input partitions highlighted in Schuz et al.'s attention analysis. In contrast, Junker et al. [24] positioned scene context as a key enhancer of model resilience, particularly under challenging conditions like occlusion. The new findings corroborate this resilience claim by demonstrating that,

under high semantic congruence, contextual cues can indeed compensate for target noise, leading to more accurate identifications. However, we also show that when the semantic fit is poor, an over-reliance on context may cause misidentification, suggesting that the benefits of contextual robustness are conditional upon its relevance. Thus, while both studies underscore the importance of context, the current work refines these insights by delineating the boundaries within which context acts as a facilitator versus when it may serve as a distractor, offering a more granular understanding of attention allocation dynamics and the use of context by VLMs.

Conclusions

In this thesis, we investigated how current vision-language models (VLMs) leverage context when generating referring expressions in real-world scenes. Motivated by longstanding findings in cognitive science that emphasize the influence of scene context on visual attention and object recognition, we introduced the *Common Objects Out-of-Context (COOCO) dataset*. *COOCO* systemically manipulates the semantic relatedness between objects and their surrounding scenes, enabling a graded assessment of how context shapes performance, particularly when direct visual cues are compromised by noise or occlusion.

Our experimental results confirmed that context functions as a support mechanism under challenging conditions. Specifically, models demonstrated higher accuracy and more semantically aligned descriptions when they could rely on supportive contextual cues, particularly in medium or heavy occlusion scenarios. However, the propensity for over-reliance on statistical co-occurrence or coarse scene features surfaced when objects presented low relatedness to the background, under these conditions, models produced scene-consistent but incorrect labels. Through in-depth attention deployment analyses, we found that, in heavily occluded instances, models shifted focus from the masked-out region to surrounding objects and global scene layouts, paralleling the top-down attentional strategies observed in human perception research.

In bridging the fields of cognitive psychology and computational modelling, this thesis makes three principal contributions:

1. **COOCO Dataset:** A systematically curated collection of scene images featuring varying degrees of semantic congruence between objects and their environments, scalable for deep learning yet rooted in psychological research on scene grammar.

2. **Cognitive-Inspired Benchmarking:** An evaluation paradigm that isolates the effect of context on object recognition by manipulating occlusion levels and semantic relatedness, providing a clear lens for investigating robustness in referential tasks.

3. **Attention-Based Insights:** Empirical evidence suggests that the attention patterns of state-of-the-art multimodal models reflect the relatedness between the target object and the scene. Moreover, when direct cues are masked, these models flexibly shift their attention to prioritize contextual information.

By exploring how these models behave under conditions of semantic violation, this thesis hopes to spark further developments in VLM design that capture the flexible context integration observed in human perception. Below, we outline the key limitations of this work and propose promising future directions.

8.6 Limitations

- **Dataset Familiarity:** The high semantic relatedness images in *COOCO* dataset are the same contained in MS COCO [31] on which is based. It is almost certain that the evaluated models were likely exposed to it during pre-training, as they were not trained from scratch. This prior exposure may have influenced their performance. Future studies should account for this factor and consider incorporating newly generated high-relatedness images into the dataset.
- **Model Size Discrepancy:** The models evaluated in this study varied in size, and some of them were quantized versions, making direct comparisons less reliable. Future work should aim to use comparable models to ensure a fair evaluation.
- **Image Quality:** In some cases, the quality of images with replaced objects was suboptimal. While we attempted to mitigate this issue through manual dataset filtering, further improvements in inpainting generation could enhance consistency.
- **Category-Specific Performance:** A more detailed analysis of the object categories that models correctly predicted would provide deeper insights into their capabilities and biases.

8.7 Future Directions

In this section, we outline several promising future directions worthy of further investigation.

8.7.1 Enhanced Input Analysis

To better understand how vision-language models utilize input information for generating, we could apply SHAP-based analysis as proposed in [42]. Their approach, MM-SHAP, provides a performance-agnostic multimodality score that quantifies the contribution of individual modalities using Shapley values. Leveraging a similar methodology could help understanding which part of the input are more used by the model to perform the REG task. Another method that may shed light on the inner reasoning of VLMs while performing REG is proposed in [75]. This approach integrates attention analysis with LLaVA-CAM. Specifically, attention scores highlight relevant regions during forward propagation, while LLaVA-CAM captures gradient changes through backward propagation, revealing key image features.

8.7.2 Refined Localization Studies

Causal mediation analysis (CMA) can determine which model components drive specific outputs by comparing a "clean" input (yielding the correct answer) with a minimally altered "corrupt" input (yielding an incorrect one). By patching hidden states between the two, one identifies the precise components that, when corrected, restore the correct prediction. Employing semantic minimal pairs—altering only one meaningful detail (e.g., a single image attribute)—ensures the input remains realistic, thereby isolating the specific layers or attention heads responsible for processing that detail [16]. The *COOCO* Dataset would be a perfect tool to use in this analysis since it is made of semantically minimally different images. Alternatively, simpler ablation studies that remove layer activations can pinpoint the processing stage at which target-related information is handled, as demonstrated in [75].

8.7.3 Syntactic Violations via Anchor Objects

A promising extension involves investigating syntactic violations by systematically rearranging the spatial positions of objects around their anchor objects in the scene. By moving items that are normally tied to these anchors into improbable or physically impossible locations, we can assess how robustly VLMs handle disruptions to conventional spatial arrangements. This approach would offer valuable insights into whether models rely primarily on learned co-occurrence statistics or can genuinely interpret hierarchical scene structure, thereby bridging knowledge from anchor-based predictions in human vision research to computational implementations of scene grammar.

Appendix A

A.1 Experiment Results Tables For All Models

In this section, we present the result tables for each evaluated model.

Rel. Level	Noise Level	Noise Area	refCLIPScore	Text-Based Similarity	Hard Acc.	Soft Acc.
low	0.000	all	0.672	0.820	0.149	0.152
low	0.000	context	0.672	0.820	0.149	0.152
low	0.000	target	0.672	0.820	0.149	0.152
low	0.500	all	0.656	0.809	0.053	0.053
low	0.500	context	0.726	0.860	0.275	0.307
low	0.500	target	0.635	0.790	0.020	0.016
low	1.000	all	0.635	0.813	0.010	0.016
low	1.000	context	0.759	0.885	0.363	0.404
low	1.000	target	0.630	0.788	0.008	0.005
middle	0.000	all	0.677	0.833	0.112	0.122
middle	0.000	context	0.677	0.833	0.112	0.122
middle	0.000	target	0.677	0.833	0.112	0.122
middle	0.500	all	0.673	0.833	0.062	0.076
middle	0.500	context	0.732	0.873	0.276	0.290
middle	0.500	target	0.657	0.817	0.036	0.037
middle	1.000	all	0.638	0.829	0.015	0.031
middle	1.000	context	0.756	0.890	0.354	0.378
middle	1.000	target	0.652	0.814	0.022	0.022
original	0.000	all	0.703	0.876	0.224	0.284
original	0.000	context	0.703	0.876	0.224	0.284
original	0.000	target	0.703	0.876	0.224	0.284
original	0.500	all	0.695	0.873	0.172	0.242
original	0.500	context	0.746	0.923	0.481	0.563
original	0.500	target	0.688	0.859	0.124	0.176
original	1.000	all	0.654	0.854	0.052	0.079
original	1.000	context	0.772	0.947	0.605	0.712
original	1.000	target	0.679	0.850	0.074	0.128

Table A.1: Results for model: Qwen/Qwen2-VL-2B-Instruct-GPTQ-Int8

Rel. Level	Noise Level	Noise Area	refCLIPScore	Text-Based Similarity	Hard Acc.	Soft Acc.
low	0.000	all	0.735	0.859	0.315	0.336
low	0.000	context	0.735	0.859	0.315	0.336
low	0.000	target	0.735	0.859	0.315	0.336
low	0.500	all	0.669	0.807	0.059	0.063
low	0.500	context	0.756	0.873	0.350	0.371
low	0.500	target	0.640	0.787	0.021	0.025
low	1.000	all	0.630	0.794	0.007	0.007
low	1.000	context	0.760	0.876	0.353	0.374
low	1.000	target	0.629	0.779	0.005	0.002
middle	0.000	all	0.749	0.881	0.339	0.368
middle	0.000	context	0.749	0.881	0.339	0.368
middle	0.000	target	0.749	0.881	0.339	0.368
middle	0.500	all	0.687	0.831	0.094	0.107
middle	0.500	context	0.763	0.889	0.353	0.392
middle	0.500	target	0.673	0.821	0.062	0.072
middle	1.000	all	0.632	0.808	0.013	0.016
middle	1.000	context	0.762	0.888	0.341	0.382
middle	1.000	target	0.655	0.808	0.021	0.020
original	0.000	all	0.776	0.950	0.660	0.779
original	0.000	context	0.776	0.950	0.660	0.779
original	0.000	target	0.776	0.950	0.660	0.779
original	0.500	all	0.720	0.890	0.326	0.387
original	0.500	context	0.786	0.960	0.721	0.839
original	0.500	target	0.721	0.887	0.314	0.373
original	1.000	all	0.651	0.839	0.055	0.066
original	1.000	context	0.784	0.955	0.699	0.816
original	1.000	target	0.695	0.860	0.174	0.209

Table A.2: Results for model: Salesforce/xgen-mm-phi3-mini-instruct-r-v1

A.2 Attention Deployment Visualization

Figure A.1 illustrates the attention deployment across the input image, organized by layers and output tokens.

Rel. Level	Noise Level	Noise Area	refCLIPScore	Text-Based Similarity	Hard Acc.	Soft Acc.
low	0.000	all	0.739	0.867	0.342	0.351
low	0.000	context	0.739	0.867	0.342	0.351
low	0.000	target	0.739	0.867	0.342	0.351
low	0.500	all	0.686	0.824	0.120	0.112
low	0.500	context	0.764	0.887	0.400	0.415
low	0.500	target	0.645	0.795	0.034	0.032
low	1.000	all	0.652	0.806	0.033	0.030
low	1.000	context	0.770	0.891	0.414	0.429
low	1.000	target	0.633	0.788	0.010	0.005
middle	0.000	all	0.751	0.886	0.354	0.384
middle	0.000	context	0.751	0.886	0.354	0.384
middle	0.000	target	0.751	0.886	0.354	0.384
middle	0.500	all	0.707	0.850	0.142	0.166
middle	0.500	context	0.769	0.900	0.401	0.437
middle	0.500	target	0.680	0.831	0.089	0.094
middle	1.000	all	0.663	0.827	0.040	0.056
middle	1.000	context	0.769	0.899	0.391	0.428
middle	1.000	target	0.659	0.817	0.028	0.027
original	0.000	all	0.777	0.956	0.651	0.789
original	0.000	context	0.777	0.956	0.651	0.789
original	0.000	target	0.777	0.956	0.651	0.789
original	0.500	all	0.746	0.919	0.404	0.518
original	0.500	context	0.792	0.972	0.723	0.877
original	0.500	target	0.727	0.896	0.294	0.393
original	1.000	all	0.685	0.861	0.118	0.171
original	1.000	context	0.791	0.969	0.701	0.858
original	1.000	target	0.700	0.869	0.151	0.223

Table A.3: Results for model: Salesforce/xgen-mm-phi3-mini-instruct-singleimg-r-v1.5

Rel. Level	Noise Level	Noise Area	refCLIPScore	Text-Based Similarity	Hard Acc.	Soft Acc.
low	0.000	all	0.766	0.887	0.421	0.445
low	0.000	context	0.766	0.887	0.421	0.445
low	0.000	target	0.766	0.887	0.421	0.445
low	0.500	all	0.694	0.826	0.139	0.132
low	0.500	context	0.771	0.887	0.393	0.427
low	0.500	target	0.648	0.797	0.034	0.032
low	1.000	all	0.654	0.806	0.036	0.035
low	1.000	context	0.771	0.887	0.386	0.424
low	1.000	target	0.634	0.789	0.007	0.004
middle	0.000	all	0.764	0.894	0.379	0.415
middle	0.000	context	0.764	0.894	0.379	0.415
middle	0.000	target	0.764	0.894	0.379	0.415
middle	0.500	all	0.708	0.847	0.141	0.165
middle	0.500	context	0.766	0.891	0.355	0.394
middle	0.500	target	0.679	0.829	0.070	0.080
middle	1.000	all	0.667	0.827	0.039	0.048
middle	1.000	context	0.765	0.888	0.332	0.374
middle	1.000	target	0.657	0.817	0.022	0.022
original	0.000	all	0.777	0.950	0.681	0.768
original	0.000	context	0.777	0.950	0.681	0.768
original	0.000	target	0.777	0.950	0.681	0.768
original	0.500	all	0.743	0.912	0.432	0.515
original	0.500	context	0.784	0.956	0.696	0.804
original	0.500	target	0.726	0.893	0.321	0.375
original	1.000	all	0.693	0.868	0.140	0.186
original	1.000	context	0.779	0.945	0.630	0.728
original	1.000	target	0.696	0.866	0.171	0.193

Table A.4: Results for model: cyan2k/molmo-7B-O-bnb-4bit

Rel. Level	Noise Level	Noise Area	refCLIPScore	Text-Based Similarity	Hard Acc.	Soft Acc.
low	0.000	all	0.750	0.879	0.373	0.380
low	0.000	context	0.750	0.879	0.373	0.380
low	0.000	target	0.750	0.879	0.373	0.380
low	0.500	all	0.677	0.818	0.108	0.094
low	0.500	context	0.751	0.877	0.357	0.372
low	0.500	target	0.654	0.804	0.041	0.040
low	1.000	all	0.645	0.805	0.025	0.023
low	1.000	context	0.754	0.878	0.352	0.381
low	1.000	target	0.639	0.796	0.008	0.005
middle	0.000	all	0.748	0.884	0.325	0.362
middle	0.000	context	0.748	0.884	0.325	0.362
middle	0.000	target	0.748	0.884	0.325	0.362
middle	0.500	all	0.696	0.842	0.112	0.133
middle	0.500	context	0.754	0.886	0.331	0.368
middle	0.500	target	0.684	0.835	0.080	0.093
middle	1.000	all	0.659	0.826	0.032	0.045
middle	1.000	context	0.753	0.883	0.315	0.350
middle	1.000	target	0.661	0.822	0.026	0.028
original	0.000	all	0.786	0.966	0.733	0.835
original	0.000	context	0.786	0.966	0.733	0.835
original	0.000	target	0.786	0.966	0.733	0.835
original	0.500	all	0.736	0.907	0.379	0.445
original	0.500	context	0.788	0.968	0.759	0.853
original	0.500	target	0.734	0.907	0.369	0.459
original	1.000	all	0.683	0.863	0.142	0.165
original	1.000	context	0.787	0.964	0.732	0.833
original	1.000	target	0.703	0.877	0.200	0.267

Table A.5: Results for model: llava-hf/llava-onevision-qwen2-0.5b-si-hf

Rel. Level	Noise Level	Noise Area	refCLIPScore	Text-Based Similarity	Hard Acc.	Soft Acc.
low	0.000	all	0.712	0.827	0.235	0.217
low	0.000	context	0.712	0.827	0.235	0.217
low	0.000	target	0.712	0.827	0.235	0.217
low	0.500	all	0.670	0.803	0.105	0.095
low	0.500	context	0.715	0.829	0.213	0.209
low	0.500	target	0.634	0.769	0.020	0.017
low	1.000	all	0.640	0.797	0.027	0.029
low	1.000	context	0.721	0.833	0.206	0.216
low	1.000	target	0.618	0.758	0.004	0.002
middle	0.000	all	0.729	0.852	0.251	0.264
middle	0.000	context	0.729	0.852	0.251	0.264
middle	0.000	target	0.729	0.852	0.251	0.264
middle	0.500	all	0.690	0.827	0.106	0.116
middle	0.500	context	0.732	0.854	0.240	0.251
middle	0.500	target	0.658	0.800	0.043	0.051
middle	1.000	all	0.654	0.818	0.031	0.040
middle	1.000	context	0.733	0.855	0.224	0.241
middle	1.000	target	0.639	0.785	0.019	0.024
original	0.000	all	0.769	0.929	0.530	0.673
original	0.000	context	0.769	0.929	0.530	0.673
original	0.000	target	0.769	0.929	0.530	0.673
original	0.500	all	0.733	0.894	0.326	0.418
original	0.500	context	0.771	0.931	0.518	0.668
original	0.500	target	0.702	0.861	0.209	0.286
original	1.000	all	0.680	0.857	0.112	0.176
original	1.000	context	0.762	0.917	0.421	0.573
original	1.000	target	0.668	0.832	0.098	0.160

Table A.6: Results for model: microsoft/kosmos-2-patch14-224

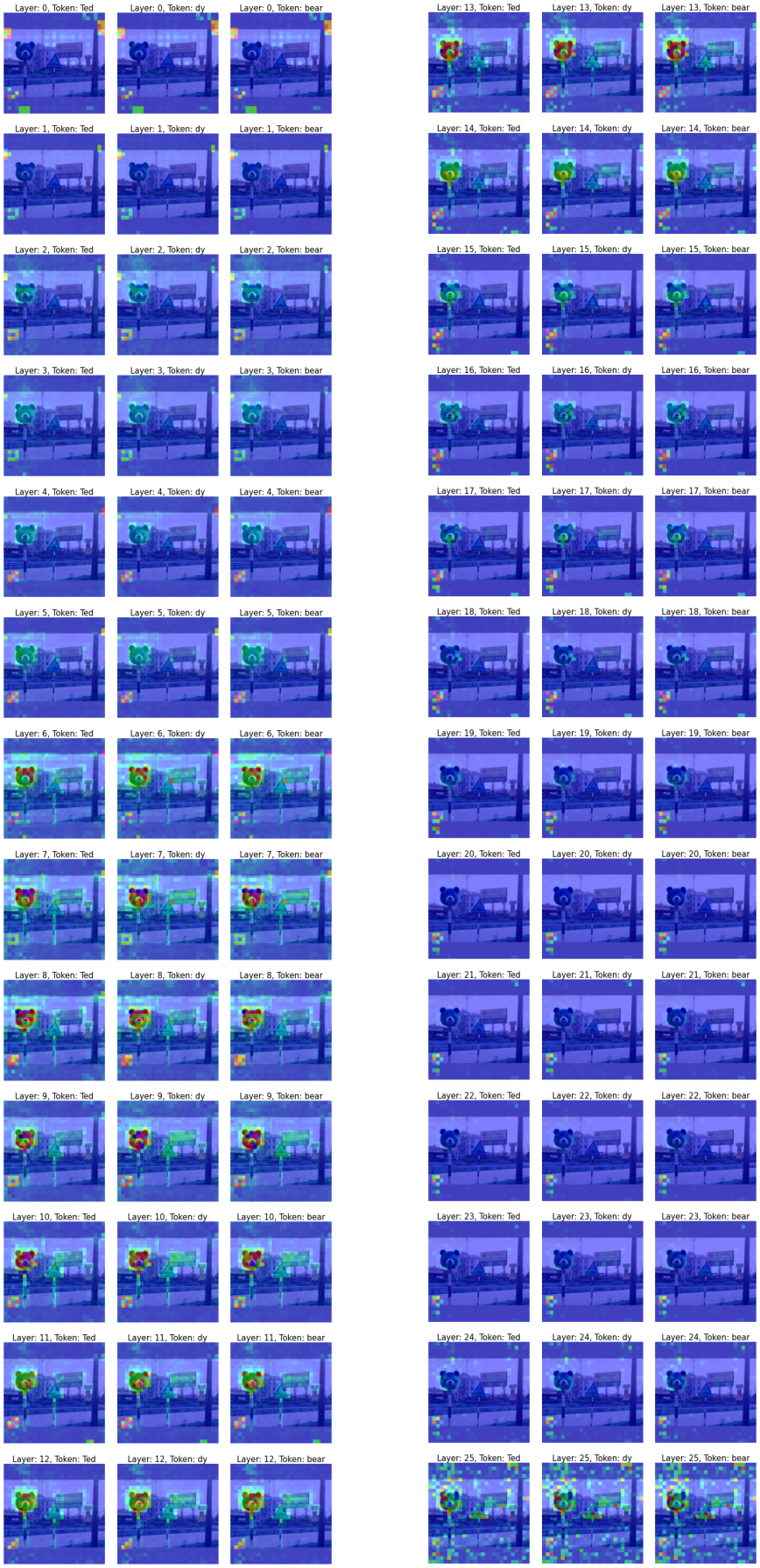


Figure A.1: Visualization of attention distribution across the input image by the LLaVA-OneVision model, analyzed per layer and per token.

Bibliography

[1] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein et al. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/ (visited on 02/21/2025).

[2] Moshe Bar. "Visual objects in context". In: *Nature Reviews Neuroscience* 5.8 (Aug. 2004), pp. 617–629. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn1476. URL: https://www.nature.com/articles/nrn1476 (visited on 06/01/2024).

[3] Sage E. P. Boettcher et al. "Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search". In: *Journal of Vision* 18.13 (Dec. 18, 2018), p. 11. ISSN: 1534-7362. DOI: 10.1167/18.13.11. URL: http://jov.arvojournals.org/article.aspx?doi=10.1167/18.13.11 (visited on 03/15/2024).

[4] Philipp Bomatter et al. *When Pigs Fly: Contextual Reasoning in Synthetic and Natural Scenes*. Aug. 11, 2021. DOI: 10.48550/arXiv.2104.02215. arXiv: 2104.02215[cs]. URL: http://arxiv.org/abs/2104.02215 (visited on 02/17/2025).

[5] Yingshan Chang et al. *WebQA: Multihop and Multimodal QA*. Mar. 28, 2022. DOI: 10.48550/arXiv.2109.00590. arXiv: 2109.00590[cs]. URL: http://arxiv.org/abs/2109.00590 (visited on 02/12/2025).

[6] Keqin Chen et al. *Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic*. July 3, 2023. arXiv: 2306.15195[cs]. URL: http://arxiv.org/abs/2306.15195 (visited on 10/15/2024).

[7] Yupei Chen et al. "Characterizing Target-absent Human Attention". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). New Orleans, LA, USA: IEEE, June 2022, pp. 5027–5036. ISBN: 978-1-66548-739-9. DOI: 10.1109/CVPRW56347.2022.00551. URL: https://ieeexplore.ieee.org/document/9857378/ (visited on 01/13/2025).

[8] Yupei Chen et al. "COCO-Search18 fixation dataset for predicting goal-directed attention control". In: *Scientific Reports* 11.1 (Apr. 22, 2021), p. 8776. ISSN: 2045-2322. DOI: 10.1038/s41598-021-87715-9. URL: https://www.nature.com/articles/s41598-021-87715-9 (visited on 03/21/2024).

[9] Claudia Damiano, Maarten Leemans, and Johan Wagemans. "Exploring the Semantic-Inconsistency Effect in Scenes Using a Continuous Measure of Linguistic-Semantic Similarity". In: *Psychological Science* 35.6 (June 2024), pp. 623–634. ISSN: 0956-7976, 1467-9280. DOI: 10.1177/09567976241238217. URL: https://journals.sagepub.com/doi/10.1177/09567976241238217 (visited on 08/01/2024).

[10] Matt Deitke et al. *Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models*. Sept. 25, 2024. arXiv: 2409.17146[cs]. URL: http://arxiv.org/abs/2409.17146 (visited on 10/02/2024).

[11] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition. ISSN: 1063-6919. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848. URL: https://ieeexplore.ieee.org/document/5206848/?arnumber=5206848 (visited on 01/14/2025).

[12] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. May 24, 2019. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.04805[cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 01/14/2025).

[13] Zhiwei Ding et al. *Efficient Zero-shot Visual Search via Target and Context-aware Transformer*. Nov. 24, 2022. DOI: 10.48550/arXiv.2211.13470. arXiv: 2211.13470[cs]. URL: http://arxiv.org/abs/2211.13470 (visited on 02/15/2025).

[14] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 3, 2021. DOI: 10.48550/arXiv.2010.11929. arXiv: 2010.11929[cs]. URL: http://arxiv.org/abs/2010.11929 (visited on 01/14/2025).

[15] Haley C. Dresang, Michael Walsh Dickey, and Tessa C. Warren. "Semantic memory for objects, actions, and events: A novel test of event-related conceptual semantic knowledge". In: *Cognitive Neuropsychology* 36.7 (Nov. 17, 2019), pp. 313–335. ISSN: 0264-3294, 1464-0627. DOI: 10.1080/02643294.2019.1656604. URL: https://www.tandfonline.com/doi/full/10.1080/02643294.2019.1656604 (visited on 01/25/2025).

[16] Michal Golovanevsky et al. *What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Noise-free Text-Image Corruption and Evaluation*. June 24, 2024. arXiv: 2406.16320[cs]. URL: http://arxiv.org/abs/2406.16320 (visited on 10/03/2024).

[17] Taylor R Hayes and John M Henderson. "Looking for Semantic Similarity: What a Vector-Space Model of Semantics Can Tell Us About Attention in Real-World Scenes". In: ().

[18] Martin N. Hebart et al. "THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images". In: *PLOS ONE* 14.10 (Oct. 15, 2019). Ed. by Fabian A. Soto, e0223792. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0223792. URL: https://dx.plos.org/10.1371/journal.pone.0223792 (visited on 01/15/2025).

[19] John M. Henderson et al. "Meaning and Attentional Guidance in Scenes: A Review of the Meaning Map Approach". In: *Vision* 3.2 (May 10, 2019), p. 19. ISSN: 2411-5150. DOI: 10.3390/vision3020019. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6802777/ (visited on 02/15/2025).

[20] Jack Hessel et al. *CLIPScore: A Reference-free Evaluation Metric for Image Captioning*. Mar. 23, 2022. arXiv: 2104.08718. URL: http://arxiv.org/abs/2104.08718 (visited on 11/13/2024).

[21] Shaohan Huang et al. *Language Is Not All You Need: Aligning Perception with Language Models*. Mar. 1, 2023. DOI: 10.48550/arXiv.2302.14045. arXiv: 2302.14045[cs]. URL: http://arxiv.org/abs/2302.14045 (visited on 02/24/2025).

[22] Benoit Jacob et al. *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference*. Dec. 15, 2017. DOI: 10.48550/arXiv.1712.05877. arXiv: 1712.05877[cs]. URL: http://arxiv.org/abs/1712.05877 (visited on 01/16/2025).

[23] Ming Jiang et al. "SALICON: Saliency in Context". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, June 2015, pp. 1072–1080. ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298710. URL: http://ieeexplore.ieee.org/document/7298710/ (visited on 01/13/2025).

[24] Simeon Junker and Sina Zarrieß. *Resilience through Scene Context in Visual Referring Expression Generation*. Apr. 18, 2024. arXiv: 2404.12289[cs]. URL: http://arxiv.org/abs/2404.12289 (visited on 04/25/2024).

[25] Emiel Krahmer and Kees Van Deemter. "Computational Generation of Referring Expressions: A Survey". In: *Computational Linguistics* 38.1 (Mar. 2012), pp. 173–218. ISSN: 0891-2017, 1530-9312. DOI: 10.1162/COLI_a_00088. URL: https://direct.mit.edu/coli/article/38/1/173-218/2136 (visited on 12/13/2024).

[26] Ranjay Krishna et al. *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*. Feb. 23, 2016. DOI: 10.48550/arXiv.1602.07332. arXiv: 1602.07332[cs]. URL: http://arxiv.org/abs/1602.07332 (visited on 01/13/2025).

[27] Tim Lauer, Filipp Schmidt, and Melissa L.-H. Võ. "The role of contextual materials in object recognition". In: *Scientific Reports* 11.1 (Nov. 9, 2021), p. 21988. ISSN: 2045-2322. DOI: 10.1038/s41598-021-01406-z. URL: https://www.nature.com/articles/s41598-021-01406-z (visited on 02/15/2025).

[28] Tim Lauer et al. "The role of scene summary statistics in object recognition". In: *Scientific Reports* 8.1 (Oct. 2, 2018), p. 14666. ISSN: 2045-2322. DOI: 10.1038/s41598-018-32991-1. URL: https://www.nature.com/articles/s41598-018-32991-1 (visited on 02/18/2025).

[29] Bo Li et al. *LLaVA-OneVision: Easy Visual Task Transfer*. Oct. 26, 2024. DOI: 10.48550/arXiv.2408.03326. arXiv: 2408.03326. URL: http://arxiv.org/abs/2408.03326 (visited on 12/03/2024).

[30] Junnan Li et al. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. Feb. 15, 2022. arXiv: 2201.12086. URL: http://arxiv.org/abs/2201.12086 (visited on 10/16/2024).

[31] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. Feb. 21, 2015. DOI: 10.48550/arXiv.1405.0312. arXiv: 1405.0312[cs]. URL: http://arxiv.org/abs/1405.0312 (visited on 01/13/2025).

[32] Haotian Liu et al. *Improved Baselines with Visual Instruction Tuning*. May 15, 2024. DOI: 10.48550/arXiv.2310.03744. arXiv: 2310.03744[cs]. URL: http://arxiv.org/abs/2310.03744 (visited on 02/25/2025).

[33] Haotian Liu et al. *Visual Instruction Tuning*. Dec. 11, 2023. DOI: 10.48550/arXiv.2304.08485. arXiv: 2304.08485[cs]. URL: http://arxiv.org/abs/2304.08485 (visited on 01/15/2025).

[34] Haotian Liu et al. *Visual Instruction Tuning*. Dec. 11, 2023. DOI: 10.48550/arXiv.2304.08485. arXiv: 2304.08485[cs]. URL: http://arxiv.org/abs/2304.08485 (visited on 01/27/2025).

[35] Junhua Mao et al. *Generation and Comprehension of Unambiguous Object Descriptions*. Apr. 11, 2016. DOI: 10.48550/arXiv.1511.02283. arXiv: 1511.02283[cs]. URL: http://arxiv.org/abs/1511.02283 (visited on 02/26/2025).

[36] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. Sept. 7, 2013. DOI: 10.48550/arXiv.1301.3781. arXiv: 1301.3781[cs]. URL: http://arxiv.org/abs/1301.3781 (visited on 01/15/2025).

[37] Emiel van Miltenburg et al. "DIDEC: The Dutch Image Description and Eye-tracking Corpus". In: *Proceedings of the 27th International Conference on Computational Linguistics*. COLING 2018. Ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3658–3669. URL: https://aclanthology.org/C18-1310/ (visited on 01/13/2025).

[38] Ron Mokady, Amir Hertz, and Amit H. Bermano. *ClipCap: CLIP Prefix for Image Captioning*. Nov. 18, 2021. DOI: 10.48550/arXiv.2111.09734. arXiv: 2111.09734[cs]. URL: http://arxiv.org/abs/2111.09734 (visited on 02/21/2025).

[39] Sabine Öhlschläger and Melissa Le-Hoa Võ. "SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes". In: *Behavior Research Methods* 49.5 (Oct. 2017), pp. 1780–1791. ISSN: 1554-3528. DOI: 10.3758/s13428-016-0820-3. URL: http://link.springer.com/10.3758/s13428-016-0820-3 (visited on 03/13/2024).

[40] Aude Oliva and Antonio Torralba. "The role of context in object recognition". In: *Trends in Cognitive Sciences* 11.12 (Dec. 2007), pp. 520–527. ISSN: 13646613. DOI: 10.1016/j.tics.2007.09.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661307002550 (visited on 02/16/2025).

[41] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. ACL 2002. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040/ (visited on 02/21/2025).

[42] Letitia Parcalabescu and Anette Frank. *MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks*. Sept. 18, 2024. DOI: 10.48550/arXiv.2212.08158. arXiv: 2212.08158. URL: http://arxiv.org/abs/2212.08158 (visited on 11/29/2024).

[43] Marius V. Peelen, Eva Berlot, and Floris P. De Lange. "Predictive processing of scenes and objects". In: *Nature Reviews Psychology* 3.1 (Nov. 23, 2023), pp. 13–26. ISSN: 2731-0574. DOI: 10.1038/s44159-023-00254-0. URL: https://www.nature.com/articles/s44159-023-00254-0 (visited on 02/15/2025).

[44] Zhiliang Peng et al. *Kosmos-2: Grounding Multimodal Large Language Models to the World*. July 13, 2023. arXiv: 2306.14824[cs]. URL: http://arxiv.org/abs/2306.14824 (visited on 09/23/2024).

[45] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2014. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/ (visited on 01/15/2025).

[46] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: ().

[47] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. Feb. 26, 2021. arXiv: 2103.00020[cs]. URL: http://arxiv.org/abs/2103.00020 (visited on 11/11/2023).

[48] Tal Ridnik et al. *ImageNet-21K Pretraining for the Masses*. Aug. 5, 2021. DOI: 10.48550/arXiv.2104.10972. arXiv: 2104.10972[cs]. URL: http://arxiv.org/abs/2104.10972 (visited on 01/14/2025).

[49] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. Apr. 13, 2022. DOI: 10.48550/arXiv.2112.10752. arXiv: 2112.10752[cs]. URL: http://arxiv.org/abs/2112.10752 (visited on 01/16/2025).

[50] Simeon Schüz, Albert Gatt, and Sina Zarrieß. "Rethinking symbolic and visual context in Referring Expression Generation". In: *Frontiers in Artificial Intelligence* 6 (Mar. 21, 2023), p. 1067125. ISSN: 2624-8212. DOI: 10.3389/frai.2023.1067125. URL: https://www.frontiersin.org/articles/10.3389/frai.2023.1067125/full (visited on 09/22/2024).

[51] Simeon Schüz and Sina Zarrieß. "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)*. Ed. by Albert Gatt et al. Prague, Czech Republic: Association for Computational Linguistics, Sept. 2023, pp. 20–27. URL: https://aclanthology.org/2023.mmnlg-1.3 (visited on 06/01/2024).

[52] Yarden Shir, Naphtali Abudarham, and Liad Mudrik. "You won't believe what this guy is doing with the potato: The ObjAct stimulus-set depicting human actions on congruent and incongruent objects". In: *Behavior Research Methods* 53.5 (Oct. 2021), pp. 1895–1909. ISSN: 1554-3528. DOI: 10.3758/s13428-021-01540-6. URL: https://link.springer.com/10.3758/s13428-021-01540-6 (visited on 03/13/2024).

[53] Eelke Spaak and Marius V Peelen. "Scene Context Impairs Perception of Semantically Congruent Objects". In: ().

[54] Robyn Speer, Joshua Chin, and Catherine Havasi. *ConceptNet 5.5: An Open Multilingual Graph of General Knowledge*. Dec. 11, 2018. DOI: 10.48550/arXiv.1612.03975. arXiv: 1612.03975[cs]. URL: http://arxiv.org/abs/1612.03975 (visited on 01/15/2025).

[55] Laura M. Stoinski, Jonas Perkuhn, and Martin N. Hebart. "THINGSplus: New norms and metadata for the THINGS database of 1854 object concepts and 26,107 natural object images". In: *Behavior Research Methods* (Apr. 24, 2023). ISSN: 1554-3528. DOI: 10.3758/s13428-023-02110-8. URL: https://link.springer.com/10.3758/s13428-023-02110-8 (visited on 03/11/2024).

[56] Roman Suvorov et al. *Resolution-robust Large Mask Inpainting with Fourier Convolutions*. Nov. 11, 2021. DOI: 10.48550/arXiv.2109.07161. arXiv: 2109.07161[cs]. URL: http://arxiv.org/abs/2109.07161 (visited on 01/15/2025).

[57] Antonio Torralba et al. "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search." In: *Psychological Review* 113.4 (Oct. 2006), pp. 766–786. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/0033-295X.113.4.766. URL: https://doi.apa.org/doi/10.1037/0033-295X.113.4.766 (visited on 03/09/2024).

[58] Ashish Vaswani et al. *Attention Is All You Need*. Dec. 5, 2017. arXiv: 1706.03762[cs]. URL: http://arxiv.org/abs/1706.03762 (visited on 05/22/2023).

[59] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. *CIDEr: Consensus-based Image Description Evaluation*. June 3, 2015. DOI: 10.48550/arXiv.1411.5726. arXiv: 1411.5726[cs]. URL: http://arxiv.org/abs/1411.5726 (visited on 02/21/2025).

[60] Jesse Vig and Yonatan Belinkov. "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*. BlackboxNLP 2019. Ed. by Tal Linzen et al. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 63–76. DOI: 10.18653/v1/W19-4808. URL: https://aclanthology.org/W19-4808/ (visited on 01/27/2025).

[61] Melissa Le-Hoa Võ. "The meaning and structure of scenes". In: *Vision Research* 181 (Apr. 2021), pp. 10–20. ISSN: 00426989. DOI: 10.1016/j.visres.2020.11.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S0042698920301796 (visited on 03/09/2024).

[62] Melissa Le-Hoa Võ, Sage Ep Boettcher, and Dejan Draschkow. "Reading scenes: how scene grammar guides attention and aids perception in real-world environments". In: *Current Opinion in Psychology* 29 (Oct. 2019), pp. 205–210. ISSN: 2352250X. DOI: 10.1016/j.copsyc.2019.03.009. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352250X18302574 (visited on 03/15/2024).

[63] Peng Wang et al. *Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution*. Oct. 3, 2024. arXiv: 2409.12191. URL: http://arxiv.org/abs/2409.12191 (visited on 10/15/2024).

[64] Bichen Wu et al. *Visual Transformers: Token-based Image Representation and Processing for Computer Vision*. Nov. 20, 2020. DOI: 10.48550/arXiv.2006.03677. arXiv: 2006.03677[cs]. URL: http://arxiv.org/abs/2006.03677 (visited on 01/14/2025).

[65] Jianxiong Xiao et al. "SUN Database: Exploring a Large Collection of Scene Categories". In: *International Journal of Computer Vision* 119.1 (Aug. 2016), pp. 3–22. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-014-0748-y. URL: http://link.springer.com/10.1007/s11263-014-0748-y (visited on 01/14/2025).

[66] Jianxiong Xiao et al. "SUN database: Large-scale scene recognition from abbey to zoo". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, CA, USA: IEEE, June 2010, pp. 3485–3492. ISBN: 978-1-4244-6984-0. DOI: 10.1109/CVPR.2010.5539970. URL: http://ieeexplore.ieee.org/document/5539970/ (visited on 01/14/2025).

[67] Le Xue et al. *xGen-MM (BLIP-3): A Family of Open Large Multimodal Models*. Aug. 28, 2024. arXiv: 2408.08872. URL: http://arxiv.org/abs/2408.08872 (visited on 10/22/2024).

[68] An Yang et al. *Qwen2 Technical Report*. Sept. 10, 2024. DOI: 10.48550/arXiv.2407.10671. arXiv: 2407.10671[cs]. URL: http://arxiv.org/abs/2407.10671 (visited on 02/01/2025).

[69] Licheng Yu et al. *Modeling Context in Referring Expressions*. Aug. 10, 2016. DOI: 10.48550/arXiv.1608.00272. arXiv: 1608.00272[cs]. URL: http://arxiv.org/abs/1608.00272 (visited on 01/13/2025).

[70] Woo-han Yun et al. "Cut-and-Paste Dataset Generation for Balancing Domain Gaps in Object Instance Detection". In: *IEEE Access* 9 (2021), pp. 14319–14329. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3051964. arXiv: 1909.11972[cs]. URL: http://arxiv.org/abs/1909.11972 (visited on 02/19/2025).

[71] Gregory J. Zelinsky et al. *Predicting Goal-directed Attention Control Using Inverse-Reinforcement Learning*. Jan. 31, 2020. DOI: 10.48550/arXiv.2001.11921. arXiv: 2001.11921[cs]. URL: http://arxiv.org/abs/2001.11921 (visited on 01/13/2025).

[72] Xiaohua Zhai et al. *Sigmoid Loss for Language Image Pre-Training*. Sept. 27, 2023. DOI: 10.48550/arXiv.2303.15343. arXiv: 2303.15343[cs]. URL: http://arxiv.org/abs/2303.15343 (visited on 01/27/2025).

[73] Hao Zhang et al. *LLaVA-Grounding: Grounded Visual Chat with Large Multimodal Models*. Dec. 5, 2023. arXiv: 2312.02949. URL: http://arxiv.org/abs/2312.02949 (visited on 10/28/2024).

[74] Shilong Zhang et al. *GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest*. June 1, 2024. arXiv: 2307.03601. URL: http://arxiv.org/abs/2307.03601 (visited on 10/15/2024).

[75] Xiaofeng Zhang et al. *From Redundancy to Relevance: Information Flow in LVLMs Across Reasoning Tasks*. Oct. 17, 2024. DOI: 10.48550/arXiv.2406.06579. arXiv: 2406.06579[cs]. URL: http://arxiv.org/abs/2406.06579 (visited on 01/27/2025).

[76] Junhao Zhuang et al. *A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting*. July 23, 2024. DOI: 10.48550/arXiv.2312.03594. arXiv: 2312.03594[cs]. URL: http://arxiv.org/abs/2312.03594 (visited on 01/15/2025).

List of Figures

List of Tables