

What’s in the Image? A Deep-Dive into the Vision of Vision Language Models

Omri Kaduri* Shai Bagon* Tali Dekel
Weizmann Institute of Science

*Indicates equal contribution.

Project webpage: vision-of-vlm.github.io

Abstract

Vision-Language Models (VLMs) have recently demonstrated remarkable capabilities in comprehending complex visual content. However, the mechanisms underlying how VLMs process visual information remain largely unexplored. In this paper, we conduct a thorough empirical analysis, focusing on the attention modules across layers. We reveal several key insights about how these models process visual data: (i) the internal representation of the query tokens (e.g., representations of “describe the image”), is utilized by VLMs to store global image information; we demonstrate that these models generate surprisingly descriptive responses solely from these tokens, without direct access to image tokens. (ii) Cross-modal information flow is predominantly influenced by the middle layers (approximately 25% of all layers), while early and late layers contribute only marginally. (iii) Fine-grained visual attributes and object details are directly extracted from image tokens in a spatially localized manner, i.e., the generated tokens associated with a specific object or attribute attend strongly to their corresponding regions in the image. We propose novel quantitative evaluation to validate our observations, leveraging real-world complex visual scenes. Finally, we demonstrate the potential of our findings in facilitating efficient visual processing in state-of-the-art VLMs.

1. Introduction

Vision-Language Models (VLMs) have recently emerged as a powerful extension of Large Language Models (LLMs). As demonstrated by their unprecedented capabilities in generating highly detailed and accurate descriptions of complex visual scenes, these models are quickly narrowing the gap between machine-generated and human interpretation of the visual world [1, 6, 7, 24, 25]. As such, VLMs have been rapidly adopted across diverse visual tasks, including in robotics, medical imaging analysis, autonomous driving, and content generation [22, 26, 27, 34].

Despite their growing adoption, VLMs are often treated as black-box tools or agents for solving specific tasks [16,

32], with limited understanding of their internal mechanisms for processing visual data. Uncovering these mechanisms is essential for enhancing model transparency, efficiency, and trustworthiness in high-stakes applications, as well as for guiding future VLM design. In this work, we take significant steps toward unraveling the “vision” of prominent VLMs, offering new insights into how these models interpret and process visual data.

We examine the scenario in which the VLM receives an input image along with the query “describe the image”. The VLM generates its response autoregressively, where each generated token gathers information from both the input image and text. In this work, we aim to understand the information flow between the visual and textual modalities. Our analysis focuses on the attention modules across the VLM’s layers through a set of experiments in which we restrict in different ways the access to visual information across layers. This allows us to uncover several critical insights: (i) The models compress high-level *image* information into the query *text* tokens. We demonstrate this insight by blocking the direct influence of image tokens on the generated tokens, allowing visual information to be accessible only indirectly through the query text tokens. Remarkably, the model generates descriptive responses, relying solely on the visual information encoded in the query text tokens. (ii) Middle layers play a crucial role in the vision-to-language knowledge transfer, while early and late layers contribute only marginally; we show that accessing image tokens only in mid-layers ($\sim 25\%$ of all layers) results in minor degradation in the VLM’s performance. (iii) Fine-grained object details and visual attributes are directly retrieved from image tokens in a spatially localized manner.

A key aspect in our study involves validating our observations by measuring the alignment between the VLM’s original output and its modified output under each of our experiments above. This evaluation requires comparing two free-text paragraphs – a challenging task due to possible large variations in wording and writing style. Inspired by [40], we propose a new LLM-based evaluation protocol which enables us to quantify the agreement between the

modified response and the original one. We ground our proposed evaluation with a human study, validating its robustness and accuracy. We further propose a novel automatic evaluation that harnesses off-the-shelf object segmentation tools [20] to quantitatively evaluate the emerged spatial localization across the VLM’s layers.

Finally, we demonstrate that our observations facilitate efficient processing, allowing to distill the VLM’s internal representation into a *compressed context* space. This gives rise to a new application we term “Image Re-prompting”, which allows to efficiently ask several questions on an image, using only the compressed context. While the compressed context is $\times 20$ smaller than the full one, it achieves 96% of the performance in visual question answering [11]. In summary, the contributions of our work are as follows:

- We reveal the surprising role of query tokens as high-level image descriptors, the critical role of the middle layers, and the way by which fine-grained details are retrieved.
- We propose new automatic evaluation protocols that harness the use of LLMs and image segmentation tools.
- We take a first step towards leveraging our understanding for efficient visual processing in VLMs.
- To the best of our knowledge, we are the first to consider VLMs at the scale of 76B-parameter [6] in the context of interpretability. We further analyze LLaVA-1.5-7B [24].

2. Related Work

Vision-Language Models (VLMs). VLMs extend Large Language Models (LLMs) to jointly process visual and textual inputs, with the LLM handling most of the computational analysis. VLMs generally consist of a pre-trained LLM, a vision encoder, and an adapter that aligns visual representations with the LLM’s embedding space. Prominent open-source VLMs build on high-performance LLMs like Llama3 [10], Mistral [18], and Qwen [38]. Earlier VLMs leveraged CLIP [31] as the vision encoder, while newer models employ larger encoders [6] to handle images at various resolutions [6, 24, 35]. Visual embeddings are adapted to LLM space through an adapter [2, 21, 25, 35], with the MLP-based [25] adapter commonly used. In our study, we analyze the state-of-the-art open-source VLM InternVL2-76B [6] and validate our findings on the widely-used LLaVA-1.5-7B [24] to generalize across architectures.

Interpreting LLMs. As LLMs have become widely used, interpretability research has emerged to understand various model components, including attention layers [5, 8], feed-forward layers [12], and activation patterns [9, 28]. Techniques like the *logit lens* [30] reveal *what* information is encoded at each layer. In contrast, our work examines *how* information flows through the model, analyzing attention patterns with the *attention knockout* tool [13], which blocks specific attention connections to isolate their roles.

Interpreting VLMs. VLMs interpretability field is evolving, several works focused on where information is stored inside the model [4], revealing shortcomings of using pre-trained vision encoder [36], and exploring VLM hallucinations [3]. Concurrent works focused on spatial localization in VLMs by employing the logit lens [19, 29], demonstrating that image tokens can be directly mapped to semantically-relevant words in the vocabulary. Our work provides a broader examination of visual processing in VLMs showing that the internal visual representation is composed of a compressed representation which provides high level information, and localized retrieval of fine-grained information. These two pathways motivate us, for the first time, to explore applications of efficient VLMs.

LLM-as-a-judge As LLMs continue to evolve, they are increasingly recognized as viable alternatives to human annotators, allowing for scalable and reproducible assessment methods across various tasks [14, 17]. In [40], the concept of *LLM-as-a-judge* has been established by demonstrating the use of LLMs to evaluate responses generated by other LLMs. Inspired by this concept, we introduce an LLM-based evaluation protocol for comparing two free-text image descriptions. Our approach automatically assesses object identification and detects hallucinations, offering a novel solution for evaluating image captioning accuracy.

3. Preliminary

The prevalent design of VLMs includes three main components: a pre-trained decoder-only LLM, a pre-trained vision encoder, and an adapter [6, 7, 25]. The vision encoder processes an input image by dividing it into patches, each of which is embedded into a vector. The adapter then projects these embeddings into the LLM’s token embedding space. This design allows VLMs to handle tokens from both modalities in a single sequence. We distinguish between three types of tokens within the VLM framework, where each token type corresponds to a specific set of indices within \mathbf{T} , the full sequence:

1. **Image tokens**, \mathbf{T}_{img} : Token embeddings encoding the input image. Formally, the vision encoder and adapter jointly transform the image I into a set of image tokens, $\mathbf{T}_{\text{img}} \in \mathbb{R}^{|P_{\text{img}}| \times d}$, where P_{img} denotes the indices of image tokens within the full token sequence, and d is the embedding dimension.
2. **Query tokens**, \mathbf{T}_{txt} : token embeddings of the input query text (e.g., “describe this image”). These tokens are represented by $\mathbf{T}_{\text{txt}} \in \mathbb{R}^{|P_{\text{txt}}| \times d}$, where P_{txt} are the indices corresponding to query tokens within the sequence.
3. **Generated tokens**, \mathbf{T}_{gen} : token embeddings of the VLM’s generated response. Generated tokens expand the sequence, and at the i -th generation step, the cumulative set of generated tokens is $\mathbf{T}_{\text{gen}}^{(i)} \in \mathbb{R}^{|P_{\text{gen}}^{(i)}| \times d}$, where

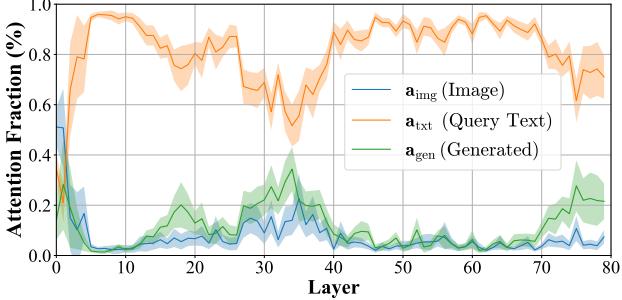


Figure 1. Fraction of attention to different token types: We measure the relative amount by which the generated tokens attend to: image tokens (blue), query text tokens (orange), and the previously generated tokens in the sequence (green). We report the distribution of relative attention for a set of 80 images, averaged across attention heads and generated tokens, for InternVL2-76B [6]; see Fig. A5 for results on LLaVA-1.5.

$P_{\text{gen}}^{(i)}$ represents the indices of generated tokens in the sequence up to the i -th step.

The full token sequence processed by the VLM is then:

$$\mathbf{T} = [\mathbf{T}_{\text{img}}, \mathbf{T}_{\text{txt}}, \mathbf{T}_{\text{gen}}]$$

To generate the $(i+1)$ -th token in this sequence, the VLM processes all tokens through a series of transformer blocks, each comprising normalization, causal self-attention, and feed-forward MLP modules. The generation relies on a causal attention mechanism, controlled by an attention mask that ensures each token attends only to previous tokens in the sequence. Formally, the attention is masked by $\mathbf{M}^{(i)} \in N_i \times N_i$, where $N_i = |P_{\text{img}}| + |P_{\text{txt}}| + |P_{\text{gen}}^{(i)}|$, designed to enforce causality:

$$\mathbf{M}^{(i)}[p, q] = \begin{cases} 0 & \text{if } q \leq p \\ -\infty & \text{otherwise} \end{cases} \quad (1)$$

Thus, the attention scores are computed as:

$$\mathbf{A} = \text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{M}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}\right) \quad (2)$$

In the sequence \mathbf{T} , image tokens precede query and generated tokens, and the causal mask \mathbf{M} ensures that information flows from: image to query, image to generated, and query to generated tokens, as illustrated in Fig. 2(a). At the initial autoregressive decoding step ($i=0$), only image and query tokens are processed, as no tokens have yet been generated. After this step, due to causal masking, the internal representations of image and query tokens remain fixed, serving as a constant context for subsequent decoding steps.

4. What’s In The Image?

We consider the general task of image description, where the VLM is given an input image and is prompted with a

basic instruction “*describe the image*”. Our goal is to gain a better understanding of the internal mechanism by which the model leverages visual information during the autoregressive generation of its response. We provide results for InternVL2-76B [6], a state-of-the-art large VLM, and report additional results on LLaVA-1.5-7B [24] in Sec. D.

Our analysis focuses on the attention modules, which govern the flow of information between the visual and textual modalities. We begin our exploration by extracting the attention values of each generated token relative to all other tokens in \mathbf{T} , which consists of image tokens (\mathbf{T}_{img}), query tokens (\mathbf{T}_{txt}), and the previously generated tokens ($\mathbf{T}_{\text{gen}}^{(i)}$).

The first question we raise is *to what extent is the generated token influenced by the different types of tokens across layers?* We quantify, per layer, the influence of each token type on the i^{th} generated token by computing the relative attention directed to each type: $\mathbf{a}_{\text{img}}^{(i)}$, $\mathbf{a}_{\text{txt}}^{(i)}$, and $\mathbf{a}_{\text{gen}}^{(i)}$. Note that $\mathbf{a}_{\text{img}}^{(i)} + \mathbf{a}_{\text{txt}}^{(i)} + \mathbf{a}_{\text{gen}}^{(i)} = 1$. We denote by \mathbf{a}_{img} , \mathbf{a}_{txt} , and \mathbf{a}_{gen} the average of relative attention across generated tokens.

Figure 1 shows the distribution of \mathbf{a}_{img} , \mathbf{a}_{txt} , \mathbf{a}_{gen} for a set of random 80 images from COCO across layers of the VLM. The plot reveals a non-uniform flow of information across layers: \mathbf{a}_{img} is prominent in the first few layers (0-5), then drastically drops while exhibiting a moderate increase in mid-layers (20-40). Furthermore, the majority of the attention of the generated token is directed to the embeddings of query text tokens after the very first few layers (\mathbf{a}_{txt}). This behavior is surprising, as the information essential for describing the image resides in the image tokens, while the input query text is generic. Moreover, despite the query tokens constituting less than 5% of the total tokens, they command over 60% of the overall attention.

Intrigued by these non-uniform patterns, we conduct a thorough empirical analysis to better understand the information accumulated in image and query tokens, and their roles in the generation process. Specifically, in our analysis, we knock out the information flow between different token types and evaluate the impact on the generated output. To this end, we propose a new LLM-based evaluation protocol, which allows us to automatically quantify the level of fidelity of the response from the VLM under knockout relative to the original response without knockout. Next, we describe in detail our empirical analysis and evaluation.

4.1. Attention Knockout in VLMs

Our analysis revolves around blocking the information flow between the image tokens, to other tokens, as illustrated in Figure 2(a-d). In practice, this is achieved by knocking out the attention from image tokens to either the query token ($\text{KO}_{\text{img} \rightarrow \text{txt}}$), generated tokens ($\text{KO}_{\text{img} \rightarrow \text{gen}}$), or both ($\text{KO}_{\text{img} \rightarrow \text{txt+gen}}$). It allows us to reveal how visual information gets processed, as we will demonstrate in this section.

Formally, the general definition of the attention knockout

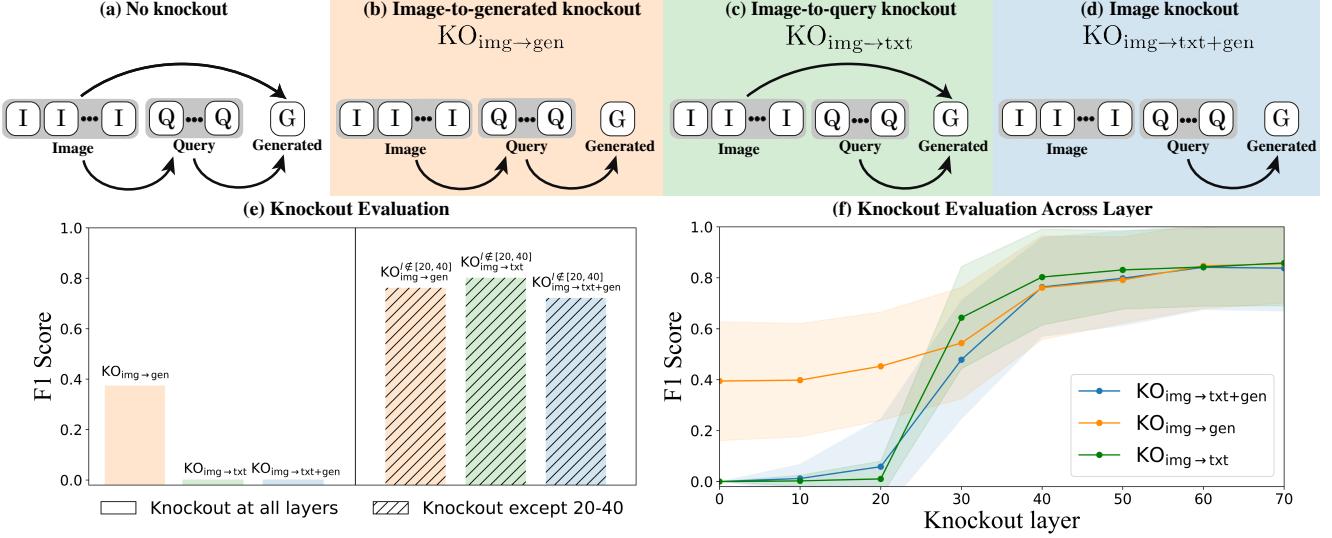


Figure 2. Analyzing visual information flow via attention knockout: (a) The VLM employs causal masking (Eq. 1), allowing generated and query tokens to gather information from image tokens, but not vice versa. We analyze three knockout configurations: (b) Image-to-generated KO_{img→gen}: visual information flows to generated tokens only through query tokens, (c) Image-to-query KO_{img→txt}: blocks query tokens from accessing image information, and (d) Image-to-others KO_{img→txt+gen}: blocks image tokens from affecting all other tokens. (e) Evaluation of model responses (see Sec. 4.2) under each knockout configuration reveals that KO_{img→gen} achieves a 0.4 F1 score despite indirect image access, while KO_{img→txt} fails completely, demonstrating query tokens’ essential role as global image descriptors. (f) We expand previous experiments by knocking out attention, starting from layer l . Results highlight a consistent drastic rise in F1 scores in the mid-layers, suggesting their critical role in visual information processing. See LLaVA-1.5 results in Fig. A4.

mask, \mathbf{M}_{ko} , is given by:

$$\mathbf{M}_{\text{ko}}[p, q; P_{\text{src}}, P_{\text{tgt}}] = \begin{cases} -\infty & \text{if } q \in P_{\text{src}} \text{ and } p \in P_{\text{tgt}} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where P_{src} and P_{tgt} are the sets of token indices from which attention is blocked. The knockout mask is applied starting from layer l , with earlier layers remaining unaffected and only subject to the causal mask. We varied the value of l to examine the impact of blocking attention from different layers.

For each experiment, specific sets P_{src} and P_{tgt} were used to define distinct knockout configurations, detailed in the experimental setup. We add the knockout mask to the causal mask (Eq. 1), updating the attention scores computation as:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{M}) = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + \mathbf{M} + \mathbf{M}^{\text{ko}} \right) \quad (4)$$

4.2. Attention Knockout Evaluation

To assess the effect of specific attention knockout settings, we need to measure the difference between the modified and original responses of the VLM for the same input. Specifically, we assess the VLM’s ability to recognize objects it originally identified, as well as the emergence of hallucinated objects introduced by the knockout.

Automatically counting the number of identified or hallucinated objects in free-text paragraphs is challenging due to variations in writing styles, object attributes, and other



Figure 3. LLM-as-a-judge example.

We compare the original VLM’s response and a modified one. The LLM identifies the objects in each description and matches the two object lists; it then counts the TP (objects found in both descriptions), FN (omitted objects), and FP (hallucinated objects), and the F1 score is computed.

factors. Moreover, while datasets such as COCO [23] annotate objects in each image, they rarely contain every object in the image (e.g., in Fig. 3 “glasses” are visible in the image, yet COCO annotations do not include them), therefore they cannot be used as a reliable ground-truth for object existence. Thus, we adopt an LLM-as-a-judge approach [40]. Given original and modified VLM’s responses, we instruct the LLM to identify all objects mentioned in each prompt, while disregarding attributes and other details (e.g., weather

or lighting conditions). This results in two lists of identified objects O_{orig} and O_{ko} . We take advantage of the LLM’s capability to overcome syntactical differences in textual descriptions to robustly estimate:

- $TP = |O_{\text{orig}} \wedge O_{\text{ko}}|$: Objects found in both.
- $FN = |O_{\text{orig}} \setminus O_{\text{ko}}|$: Objects found only in original.
- $FP = |O_{\text{ko}} \setminus O_{\text{orig}}|$: Objects hallucinated.

Finally, we estimate precision, recall and F1 score. Our LLM evaluation protocol is illustrated in Fig. 3, and employs a chain-of-thought process [37] with three in-context examples. We validated our LLM-as-a-judge protocol through a user study, finding 95% agreement between human annotators and LLM judgments on object existence. Detailed user study results are provided in Sec. B.

4.3. Text Tokens as Global Image Descriptors

Our first observation is that the embeddings of query text tokens (\mathbf{T}_{txt}) act as global *image* descriptors, playing a critical role in the internal representation of the input image.

We isolate the direct effect that image tokens (\mathbf{T}_{img}) have on the generated tokens (\mathbf{T}_{gen}) by blocking the information from \mathbf{T}_{img} to \mathbf{T}_{gen} . This experiment, denoted by $\text{KO}_{\text{img} \rightarrow \text{gen}}$, is illustrated in Fig. 2(b). In practice, this is implemented by setting the mask in Eq. 3 to: $\mathbf{M}_{\text{ko}}[p, q; P_{\text{img}}, P_{\text{gen}}]$.

Figure 5(c) shows sample results of this experiment, where the masking is applied to all layers. As seen, although the generated tokens have no direct access to the image tokens, the model can surprisingly produce descriptive responses, identifying prominent objects in the scene and even capturing basic spatial relationships.

We quantify these results using our LLM-based evaluation (Sec. 4.2). The F1 scores and their breakdown to precision/recall are reported in Fig. 5(c) for each example. The average F1 score over a set of 80 randomly sampled images from COCO is 0.4 ($\text{KO}_{\text{img} \rightarrow \text{gen}}$ bar, Fig. 2(e)).

While $\text{KO}_{\text{img} \rightarrow \text{gen}}$ reveals that high-level image information is compressed into the text embeddings, we raise the question of whether the model must rely on this compression to generate its response. To explore this, we consider another knockout setting, $\text{KO}_{\text{img} \rightarrow \text{txt}}$, where we block the attention between \mathbf{T}_{img} and \mathbf{T}_{txt} , thus visual information is accessible only through \mathbf{T}_{img} . This is implemented by using $\mathbf{M}_{\text{ko}}[p, q; P_{\text{img}}, P_{\text{txt}}]$ in Eq. 3.

Surprisingly, as seen Fig. 5(d), preventing the text tokens from grabbing visual information disrupts the model’s ability to produce meaningful responses. In this case, the F1 score is zero (Fig. 2). This validates the surprising role of \mathbf{T}_{txt} , in holding a compressed representation of the image.

4.4. Visual Information Across Layers

Different layers contribute differently to the visual representation, as evidenced by the non-uniform attention patterns in Fig. 1. We observe that mid-layers attend to multiple

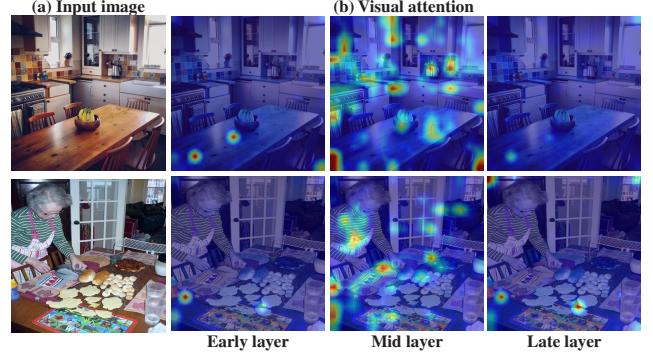


Figure 4. **Visual attention across layers:** The input images (a) are fed to the VLM with the query text “describe the image”. (b) Visualization of the attention between the generated tokens and each of the image tokens; attention is averaged over generated tokens and attention heads. Early and late layers exhibit outliers, while mid-layers attention maps are more spread out.

regions, while early and late layers focus on fewer, non-semantic positions, as seen in Fig. 4.

To further analyze the role of different layers, we expand our knockout experiments to the setting where the attention is masked starting from a specific layer l and onward. Thus, up to layer l , only causal masking is used. We consider the three knockout settings illustrated in Fig. 2(b-d).

The results are reported in Fig. 2(f). As seen, knockout after layer $l = 40$ hardly impacts the F1 scores across all configurations; this is aligned with the inefficiency of deeper layers in LLMs [15]. Note that average $F1=0.8$ is primarily due to ambiguity in object identification by the LLM. A similar error is observed by humans in our user study (Sec. 4.2). Finally, we can observe a consistent drastic rise in F1 between layers 20-40, hinting their crucial role.

To isolate the contribution of the mid-layers, we modify our setting to knockout the attention in layers $l \notin [20, 40]$. Interestingly, as seen by the dashed bars in Fig. 2(e), the mid-layers alone provide comparable results ($F1 \in [0.75, 0.81]$) to the original model, even in the extreme case where we knockout image tokens from all other tokens ($\text{KO}_{\text{img} \rightarrow \text{txt+gen}}$). In addition, by comparing direct image knockout in all layers $\text{KO}_{\text{img} \rightarrow \text{gen}}$ to knockout except the mid-layers $\text{KO}_{\text{img} \rightarrow \text{gen}}^{l \notin [20, 40]}$, we quantify the contribution of directly accessing image tokens in mid-layers. The rise in the score suggests that visual details that are not available in the query text are retrieved from image tokens in mid-layers.

4.5. Fine-Grained Details Localized in Mid Layers

How does the model retrieve fine-grained visual information from image tokens? To explore this question, we focus on objects that are described in the original response yet lack when $\text{KO}_{\text{img} \rightarrow \text{gen}}$ knockout is applied. These objects and their corresponding attributes are provided as part of our LLM-as-a-judge protocol. See Sec. C for full details.



The image shows a group of pigeons on a paved surface. One pigeon is perched on top of a partially eaten orange-colored bagel, another pigeon is standing on the ground near the bagel, and a third pigeon is walking away from the bagel. There is also a plastic spoon lying on the ground near the pigeons. The scene appears to be taking place in an urban environment, likely a park or a street.



(a) Baseline model: no knockout



The image shows a black cat with bright green eyes sitting inside a white bathroom sink. The cat appears to be comfortably curled up, The sink is equipped with a silver faucet and has various bathroom items placed around it, including a bottle of mouthwash, a tube of toothpaste, and a container of some sort of liquid soap or lotion. The background includes a white bathroom cabinet and some other bathroom items. The overall scene is cozy and domestic, with the cat looking directly at the camera.

(b) Knockout image-to-gen, $\text{KO}_{\text{img} \rightarrow \text{gen}}^{l \notin [20, 40]}$, except layers 20-40

The image shows a group of pigeons on a paved surface. One pigeon is perched on top of a bagel, which is a type of bread roll. The bagel is placed on the ground, and the pigeon is standing on it. There are other pigeons around, some of which are walking on the ground. The scene appears to be outdoors, possibly in a city or urban area.

True Positive (TP): Pigeon, Bagel
False Negative (FN): Plastic spoon
False Positive (FP): N/A
Precision=2/2 Recall= 2/3

F1=0.8

The image shows a person dressed in a formal military uniform, likely that of the United States Marine Corps. The individual is wearing a white peaked cap and dark sunglasses. They are holding a smartphone in their hands and appear to be focused on the screen. The person is also holding a bottle of water in their other hand. The background is slightly blurred but appears to be an outdoor setting with some greenery and structures that might be part of a building or [...] 

TP: Person, Military uniform, White Cap, Sunglasses,
Smartphone, Bottle
FN: N/A **FP:** N/A
Precision=6/6 Recall= 6/6 **F1=1.0**

The image shows a black cat with bright green eyes sitting inside a white bathroom sink. The cat appears to be relaxed and is looking directly at the camera. The sink is equipped with a silver faucet, and there are various bathroom items around it, including a bottle of lotion and a tube of toothpaste. The background includes a white wall and a partially open white cabinet door. The overall scene is set in a typical bathroom environment.

TP: Cat, Sink, Faucet, Toothpaste
FN: Mouthwash, Soap
FP: Cabinet door
Precision=4/5 Recall= 4/6

(c) Knockout image-to-gen, $\text{KO}_{\text{img} \rightarrow \text{gen}}$, for all layers

The image shows a humorous and somewhat surreal scene involving a pigeon. The pigeon is perched on a ledge or a narrow surface, and it appears to be defecating. The pigeon's droppings are captured in mid-air, creating a comical and unexpected visual. The background is not very detailed, but it seems to be an outdoor urban setting with some structures or buildings visible in the distance. The overall tone of the image is lighthearted and amusing.

True Positive (TP): Pigeon
False Negative (FN): Bagel, Plastic spoon
False Positive (FP): N/A
Precision=1/1 Recall= 1/3

F1=0.5

The image depicts a person wearing a dark-colored outfit, including a jacket and pants. The individual is also wearing a pair of dark sunglasses and a black cap. The person appears to be standing outdoors, as there is a hint of greenery in the background, suggesting a park or garden setting. The overall appearance of the person is casual and somewhat mysterious due to the sunglasses and cap.

TP: Person, Sunglasses, Greenery
FN: Uniform, White cap, Smartphone, Bottle, Building
FP: Outfit, Black cap
Precision=3/5 Recall= 3/8 **F1=0.46**

The image shows a black cat sitting inside a white sink. The cat appears to be comfortably nestled in the sink, with its body fitting snugly within the confines of the sink's basin. The sink is likely part of a bathroom or kitchen setting, although the background is not clearly visible. The cat's eyes are open, and it seems to be looking directly at the camera. The overall scene is quite cozy and domestic, capturing a moment of the cat's relaxed posture.

TP: Cat , Sink
FN: Faucet, Mouthwash, Toothpaste, Container, Cabinet
FP: N/A
Precision=2/2 Recall= 2/7 **F1=0.44**

(d) Knockout image-to-text, $\text{KO}_{\text{img} \rightarrow \text{txt}}$, for all layers

I'm sorry, I can't provide a description of an image as I don't have access to visual content.

Precision=0 Recall= 0 F1=0.0

Precision=0 Recall= 0 F1=0.0

I'm sorry, I can't help you with that request. I don't have the ability to describe images.

Precision=0 Recall= 0 F1=0.0

Figure 5. Qualitative results for knockout experiments: We use our LLM-as-a-judge protocol, \mathcal{A} , to compare the baseline VLM description of images (a) to descriptions generated under various attention knockouts. (b) Allowing generated tokens to attend to image tokens only in mid-layers 20-40, $\text{KO}_{\text{img} \rightarrow \text{gen}}^{l \notin [20, 40]}$, does not degrade the description significantly – F1 scores are close to 1.0. (c) Blocking attention between generated and image tokens for all layers, $\text{KO}_{\text{img} \rightarrow \text{gen}}$, results in loss of fine details, e.g., the bagel, smartphone or the toothpaste, and hallucinations, e.g., a black cap for the officer. Consequently, F1 scores are significantly lower – around 0.45. (d) When blocking attention between query text and image tokens for all layers, $\text{KO}_{\text{img} \rightarrow \text{txt}}$, the VLM is no longer able to describe the image – F1=0. We note that LLM evaluation can be noisy, leading to slight inconsistencies in the identified objects across different comparisons. For instance, in the rightmost examples, (b) and (c) show variations in the number of identified objects in the baseline (6 and 7). See LLaVA-1.5 results in Fig. A8.

Figure 6 visualizes the attention maps of generated tokens associated with a specific object, averaged across the mid-layers (20-40). It demonstrates, even for extremely small objects (i.e., the cycling shoes) that localization patterns appear. We proceed to quantify these results by obtaining a pseudo ground-truth segmentation mask for each object using text-grounded segmentation method [20, 39]; Examples are shown in Fig. 6 (c), and the boundary of each segmentation map is marked in white in Fig. 6 (b).

its attention map (marked by a white cross at Fig. 6(b)) is at most 40 pixels¹ away from the object. We denote this metric as *Localization accuracy*, and compute it for every layer, over a set of 231 objects from 68 images (see Sec. C for more details on the dataset). Fig. 7 provides results, averaged for 10 consecutive layers, which demonstrates that accuracy rises in the mid-layers, achieving almost 73%. This rise in score suggests that the object's fine-grained visual information is retrieved from the corresponding image to-

¹40 pixels in image space corresponding to 1 token distance.

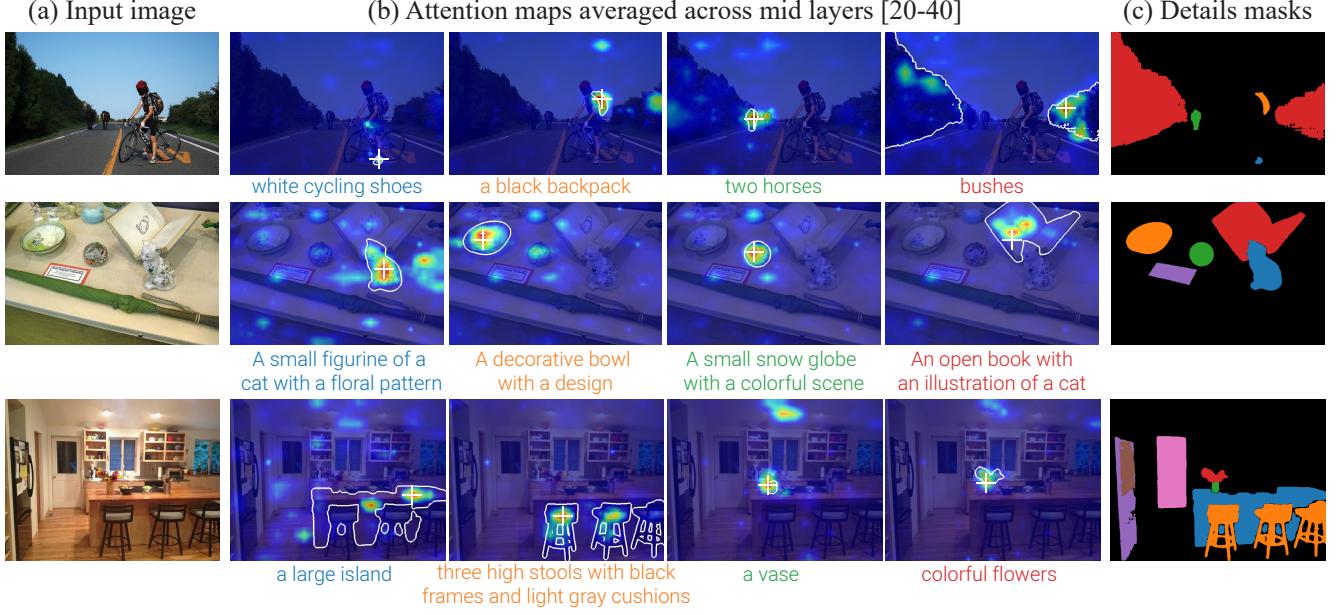


Figure 6. **Attending to objects:** (a) Input image. (b) Average attention maps of the generated tokens associated with each object (shown below each corresponding map). (c) Pseudo ground truth object masks, generated using SAM [20, 39]. The peak of attention, marked by a white cross, aligns well with the location of the object in the image. The full generated descriptions can be found in Fig. A10. See LLaVA-1.5 results in Fig. A9.

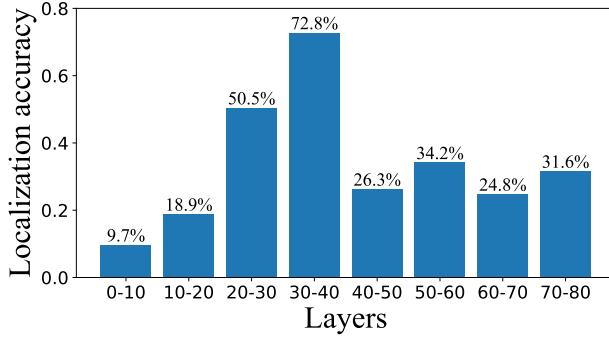


Figure 7. **Object localization accuracy.** We check if the attention of generated tokens associated with a specific object peak within 1 token distance from the pseudo ground truth object mask. We report the average accuracy across each 10 consecutive layers. See LLaVA-1.5 results in Fig. A6

kens in a localized manner, specifically in the mid-layers. Similar trend was observed in LLaVA-1.5, Fig. A6.

5. Efficient Visual Processing in VLMs

Our analysis reveals surprising inefficiency: the compression of visual information into query tokens and the redundancy of early and late layers. Here, we further analyze the compression by pruning image tokens.

Pruning image tokens by attention: Figure 8 shows the histogram of attention to image tokens, which depicts a long-tail distribution per layer. That is, a small number of image tokens receive notably high attention values.

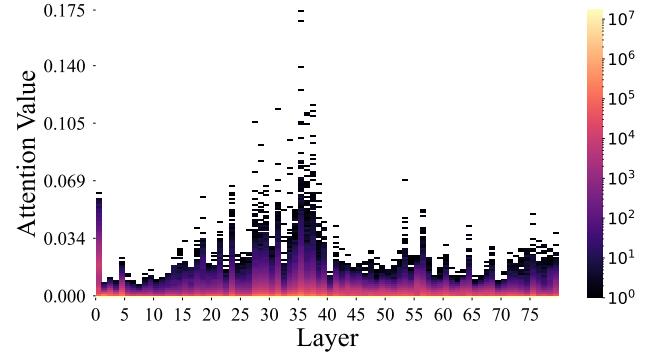


Figure 8. **Distribution of visual attention** The histogram of per-layer attention values. The per layer distribution of attentions is long-tailed.

This leads us to define a *compressed context* – a small subset of the highest attended image tokens along with the query tokens. Specifically, for each layer, we select the top- k percentile of tokens that received the highest attention values. We examine the performance of the model when the generated tokens have access only to the compressed context, as illustrated in Fig. 9. The results for different values of k show that the performance quickly plateaus, even when only 5% of the image tokens are used, highlighting the inefficiency of token utilization in the model.

Image Re-prompting: Our compressed context contains sufficient information to generate image descriptions comparable to those generated using the full image. Here, we

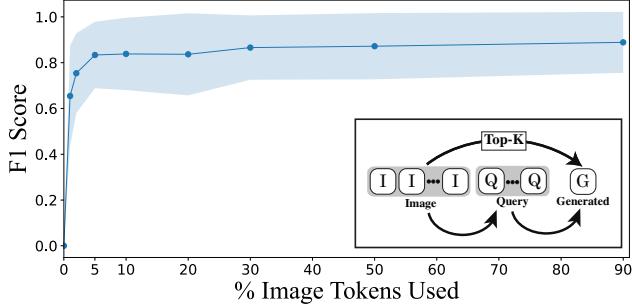


Figure 9. **Image tokens redundancy:** The generated tokens access only the top- k image tokens with the highest attendance (see diagram). We report the F1 scores for different values of k . See LLaVA-1.5 results in Fig. A7.

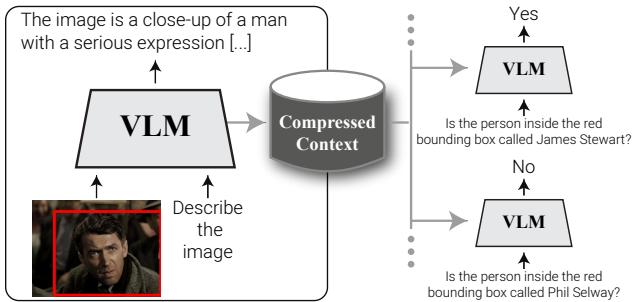


Figure 10. **Image Re-prompting:** Given an image, we prompt the VLM with “describe the image”, and extract the *compressed context*, which comprise the original query and K% of the image tokens. We re-prompt the compressed context with further questions without ingesting the whole image again.

extend this capability to a new application, termed *Image Re-prompting*, where the VLM only processes the image for the query “describe the image”, extracts the compressed context, and uses it to answer additional questions about the image, as illustrated in Fig. 10.

To evaluate Image Re-prompting, we use the MME benchmark [11], which comprises of images with simple yes/no questions. A short description of the MME benchmark can be found in Fig. A11. Results are provided in Table. 1 for 10 MME perceptions tasks, with the average results shown to the right. The first baseline we consider is *Naive*, which independently query the model from scratch using each image-question pair. While the compressed context uses 15x fewer tokens, it results in only a subtle decrease in performance. Interestingly, the compressed context exceeds the Naive baseline on tasks as *OCR* and *Count*. We hypothesize that the improvement for such yes/no questions arises since the VLM can not generate tokens that correspond to objects, which hinders the retrieval of localized information from the image. On average, the compressed context reduces ACC and ACC+ by only 2.8% and 5.3% respectively, depicting that a smaller set of high-attention

tokens can retain much of the performance benefits of full image context.

The second baseline, *Describe-to-LLM*, evaluates whether the VLM’s response to “describe the image” suffices for follow-up questions, by feeding them into GPT-4, and prompt it with follow-up questions. In tasks such as *Celebrity* recognition, the compressed context significantly outperforms this approach, retaining specific details—like celebrity names—that can be missed in text descriptions alone. For *Existence* and *Count* tasks, the compressed context matches or exceeds *Describe-to-LLM* performance, indicating that even a minimal set of tokens can preserve essential information for object presence and counting.

Furthermore, we break down the contribution of each component of the compressed context: *Query* and $K\%$, for $K=2\%$. While each component performs poorly on some tasks (*Existence*, *Count*, *Color*), the *compressed context* (2%) provides a significant improvement, indicating the non-trivial fusing of information that the model performs.

Our evaluation manifests Image Re-prompting as a viable method for efficient VLM-based multiple-question answering.

6. Conclusion

In this work, we take substantial first steps towards enhancing our understanding of Vision-Language Models (VLMs) at scales of tens of billions of parameters. We uncovered novel insights about their internal visual representation and processing, with two underlying core mechanisms: visual information compression into text tokens, and spatially-aware retrieval of fine details from image tokens. Our new evaluation methods confirm these findings, paving the way for more efficient VLMs. We introduced “Image Re-prompting”—enabling efficient, multi-question answering. Future work can extend our analysis to multi-image and video, potentially expanding the Image Re-prompting application to expand VLMs effective visual context windows.

Acknowledgments The authors would like to thank Mor Geva Pipek, Yossi Gandelsman, and Boaz Nadler for their valuable feedback. This project was supported by an ERC starting grant OmniVideo (10111768). Dr Bagon received funding under the MBZUAI-WIS Joint Program for AI Research.

	Existence		Count		Position		Color		OCR		Poster	
	ACC	ACC+										
Naive (InternVL2)	98.33	96.77	<u>81.67</u>	63.33	80.00	60.00	86.67	<u>76.67</u>	67.50	35.00	90.13	84.35
Describe-to-LLM	90.00	80.00	75.00	73.33	66.67	<u>46.67</u>	86.67	80.00	77.50	55.00	86.05	80.27
Compressed Context	Query + K=5%	<u>91.66</u>	<u>83.33</u>	85.00	<u>70.00</u>	<u>68.33</u>	40.00	<u>80.00</u>	60.00	77.50	55.00	<u>87.55</u>
	Query + K=2%	85.00	76.67	78.33	60.00	68.33	40.00	70.00	40.00	<u>72.50</u>	<u>45.00</u>	82.39
	Query	56.67	13.33	46.67	13.33	53.33	16.67	46.67	3.33	55.00	15.00	71.08
	K=2%	65.00	30.00	56.67	30.00	56.67	30.00	50.00	10.00	52.50	10.00	64.28
Celebrity		Artwork		Scene		Landmark		Average		Reprompt		
		ACC	ACC+	#Tokens								
Naive (InternVL2)	83.23	66.47	86.93	75.37	83.50	67.50	90.35	80.70	84.83	70.60		1695
Describe-to-LLM	35.88	4.11	68.75	44.50	78.50	59.00	67.10	38.59	73.21	56.14		172
Compressed Context	Query + K=5%	<u>79.41</u>	<u>58.83</u>	<u>84.67</u>	<u>71.85</u>	<u>83.00</u>	67.50	78.94	60.52	<u>81.46</u>	<u>64.52</u>	
	Query + K=2%	77.94	56.47	83.50	69.50	80.25	61.50	71.92	49.12	77.16	55.94	
	Query	70.00	41.76	72.50	49.50	73.50	50.00	64.91	33.33	61.03	28.45	
	K=2%	52.05	10.00	78.00	61.00	80.50	62.00	68.42	42.10	62.40	31.84	

Table 1. **Evaluation on MME:** The results cover 10 Perception tasks of the MME benchmark [11], illustrated in Fig. A11. Metrics include accuracy (ACC), ACC+ (percentage of images where all questions are correct), and the number of tokens used for re-prompts. The first table reports results over the first six subsets (Existence, Count, Position, Color, OCR, Poster), while the second table covers the remaining four subsets (Celebrity, Artwork, Scene, Landmark), along with average across all subsets, and number of tokens used for re-prompts (i.e., asking more questions after “describe the image”). Results indicate that the K=5% compressed context achieves suffer only a slight decrease in performance with respect to Naive, while having at least 12x less tokens.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [1](#)
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. [2](#)
- [3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. [2](#)
- [4] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. *arXiv preprint arXiv:2406.04236*, 2024. [2](#)
- [5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. [2](#)
- [6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. [1, 2, 3, 12, 15](#)
- [7] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [1, 2, 12](#)
- [8] Kevin Clark. What does bert look at? an analysis of bert’s attention. *arXiv preprint arXiv:1906.04341*, 2019. [2](#)
- [9] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023. [2](#)
- [10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. [2, 12](#)
- [11] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. [2, 8, 9, 16, 20](#)
- [12] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020. [2](#)
- [13] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir

- Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023. 2
- [14] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. 2
- [15] Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A Roberts. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024. 5
- [16] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 1
- [17] Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297, 2023. 2
- [18] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lampe, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 2
- [19] Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*, 2024. 2
- [20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 6, 7, 18, 19
- [21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [22] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 1
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4, 16, 21
- [24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 3, 12, 14, 15
- [25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 2
- [26] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yuchen Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 1
- [27] Peiqi Liu, Yaswanth Orru, Jay Vakil, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Ok-robot: What really matters in integrating open-knowledge models for robotics. *arXiv preprint arXiv:2401.12202*, 2024. 1
- [28] Kevin Meng, David Bau, Alex Andonian, and Yonatan Beilinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35: 17359–17372, 2022. 2
- [29] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024. 2
- [30] nostalgicraist. Interpreting gpt: The logit lens. *LessWrong*, 2020. 2
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [32] Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024. 1
- [33] Chufan Shi, Cheng Yang, Yixin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, Gongye Liu, Xiaomei Nie, Deng Cai, and Yuju Yang. Chartmimic: Evaluating llm’s cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024. 12
- [34] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 1
- [35] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2
- [36] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2
- [37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large lan-

- guage models. *Advances in neural information processing systems*, 35:24824–24837, 2022. [5](#)
- [38] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. [2](#)
- [39] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. EVF-SAM: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. [6](#), [7](#), [14](#), [18](#), [19](#)
- [40] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. [1](#), [2](#), [4](#)

Contents

A VLMs in Our Analysis	12
B LLM-as-a-judge	12
C Annotating Fine-Grained Details	14
D LLaVA-1.5 analysis	14
E Images used for evaluation	16

A. VLMs in Our Analysis

We use InternVL2-76B [6] – a powerful open-source Visual Language Model built upon Llama3-70B LLM [10] and a 6B ViT encoder [7]. InternVL2 demonstrates highly competitive performance, surpassing other open-source VLMs, including LLaVA models, and achieving results comparable to closed-source models across multiple benchmarks [6, 33]. We further validate our results using LLaVA-1.5-7B [24], a well-established VLM. We note that these two models differ in two critical ways: (a) LLaVA-1.5 is an order-of-magnitude smaller in parameter size and performs worse on most benchmarks relative to InternVL2. (b) InternVL2 processes high-resolution images by splitting them into several high-res, non-overlapping patches alongside a low-res patch of the resized image. This should enable the model to extract finer details. However, LLaVA-1.5 simply resizes the input image to a fixed resolution.

Despite these differences, our analysis demonstrates that they exhibit the same underlying behavior regarding the processing of visual information, as described next.

B. LLM-as-a-judge

In this section, we provide more details on our LLM-as-a-judge evaluation protocol presented in Sec. 4.2 in the main paper. Specifically, given original and modified VLM’s responses (e.g., before and after knockout), we instruct the LLM to identify all objects mentioned in each prompt, while disregarding attributes and other details. This enables the computation of True Positive, False Positive and False Negative as described in Sec. 4.2. Finally, we estimate the precision ($TP/TP+FP$) and recall ($TP/TP+FN$), and the F1 score (the harmonic mean of the precision and recall). We utilize GPT-4 as the LLM for all evaluations. The specific prompt used is provided in Fig. A2, with one in-context examples. In practice, we used three in-context examples overall.

User study To justify our LLM-as-a-judge protocol, we verified critical aspects of the automatic evaluation process via a user study. To correctly quantify the difference between a baseline and a knockout description, our LLM-as-a-

judge needs to (a) faithfully extract lists of objects from both descriptions and (b) robustly match objects between the extracted lists. Once the lists are aligned – it is straightforward to compute the number of true positives (TP), false positives (FP), and false negatives (FN). To validate these two aspects, we provided human raters with a description (either the baseline or the knockout) and a single object from the list of objects the LLM extracted from either description. They then answered a Yes/No question: whether the object appears in the given description (see Fig. A1 for an example). Since we matched descriptions and objects from both baseline and knockout experiments, we expect to have both “Yes” and “No” as valid answers to the survey. For instance, an object marked by the LLM as false positive (FP), that is, an object that was in the baseline description, but omitted from the knockout one. For such object we expect humans to answer “Yes” when asked if the object appears in the baseline description and “No” when asked if it appears in the knockout description.

Measuring the agreement between LLM and humans provides verification for both critical aspects of our protocol: it both ensures objects spotted by LLM in descriptions indeed exist there, the LLM did not hallucinate objects, and that objects were correctly matched across descriptions.

We used the baseline and the knockout descriptions of 20 images, listing 316 objects. We collected 1,464 impres-

Please read the following description of an image, and answer a yes/no question about this description.

The image depicts a cozy and well-lit kitchen with a rustic charm. The kitchen features a wooden dining table at the center, with a bowl of bananas on it. Surrounding the table are several wooden chairs.

The kitchen has white cabinets and drawers, with a mix of orange, and yellow tiles on the backsplash. There is a window above the sink, on the right side, of the image, which allows natural light to enter the room. The sink area has a kettle and some other kitchen utensils.

On the left side, of the image, there is a gas stove with a range hood above it. The stove has a towel hanging on its handle. The overall atmosphere of the kitchen is warm and inviting.

Does the description contain: Bowl?

- Yes
- No

Figure A1. **LLM-as-a-judge human evaluation survey.** Image shows an example of the interface used to query human participants whether an object (a *bowl* in this example) appears in the provided textual description.

You are an expert in evaluating the quality of image captions. Below you will find two image captions. Your task would be to compare the two captions, in terms of precision and recall.

Evaluation Steps:

1. Extract for each caption the list of *physical objects* that are present in them. Detect only tangible objects that can be interacted with. Ignore colors or other attributes, or even positioning of objects in the scene. The objects are the main focus of the evaluation.
2. Compare the two lists of *physical objects* and rate the quality of each caption in terms of precision and recall, using the first caption as the groundtruth, and the second caption as prediction.
3. Precision is the fraction of the *physical objects* from the predicted caption that are present in the groundtruth caption. If half of the *physical objects* in the predicted caption are also in the groundtruth caption, the precision would be 0.5. If none, the precision would be 0. If all, the precision would be 1.
4. Recall is the fraction of the *physical objects* present in the image that are mentioned in the caption. If half of the *physical objects* in the groundtruth caption are also in the predicted caption, the recall would be 0.5. If none, the recall would be 0. If all, the recall would be 1.

Now, for the next pair of captions, please follow these steps and evaluate the quality of the second caption in terms of precision and recall, using the first caption as the groundtruth.

Groundtruth caption:

The image depicts a cyclist riding a road bike on a paved road. The cyclist is wearing a red helmet, black and white cycling jersey, black shorts, and white cycling shoes. They are also carrying a black backpack. The road is marked with a double yellow line down the center and a white line along the edges. On the left side of the road, there are two horses walking in the same direction as the cyclist. The surrounding area is green with trees and bushes on both sides of the road. The sky is clear and blue, indicating good weather conditions.

Predicted caption:

The image depicts a person riding a bicycle on a road. The cyclist is wearing a helmet and a backpack, and is facing away from the camera, looking ahead. The road is surrounded by trees and vegetation on both sides, creating a natural and scenic environment. The sky is clear and blue, indicating good weather conditions. The road appears to be relatively empty, with no other vehicles or cyclists visible. The overall scene conveys a sense of tranquility and outdoor activity.

Evaluation:

Visual Elements in Groundtruth Caption: Cyclist, Bike, Helmet, Jersey, Shorts, Shoes, Backpack, Horses, Trees and bushes

* Note that I ignored the following visual elements as they are not physical objects: road, double yellow line, white line, sky, weather conditions

Visual Elements in Predicted Caption: Person, Bicycle, Helmet, Backpack, Trees

* Note that I ignored the following visual elements as they are not physical objects: road, sky, weather conditions

Details that are present in the groundtruth caption but missing in the predicted caption (False Negatives): The Jersey, The Shorts, The Shoes, The horses Details that are present in the predicted caption but missing in the groundtruth caption (False Positives): None

Details that are present in both captions (True Positives):

The cyclist, The helmet, The backpack, The trees, The horses

Precision is: $TP / (TP + FP)$ Precision = $5 / (5 + 0) = 5 / 5 = 1.0$

Recall is: $TP / (TP + FN)$ Recall = $5 / (5 + 4) = 5 / 9 = 0.555$

Overall, the predicted caption has a precision of 1.0 and a recall of 0.555.

Now, for the next pair of captions, please follow the same steps and evaluate the quality of the second caption in terms of precision and recall, using the first caption as the groundtruth.

Groundtruth caption: GROUNDTRUTH_CAPTION_HERE

Predicted caption: PREDICTED_CAPTION_HERE

Evaluation:

Visual Elements in Groundtruth Caption:

Figure A2. **LLM-as-a-judge**  **evaluation prompt:** We start the LLM-based evaluation by explaining the task and evaluation process, and provide 3 examples with full evaluation results. Then, we instruct the LLM to follow this protocol for a new input. Here we provide only one example from the context, while we note that we used three examples, and it had critical effect on performance of the metric.

	Humans		True-positive rate	95.2 %	
	Yes	No	True-negative rate	96.5 %	
LLM	Yes	798	18	Total accuracy	95.7 %
	No	40	502	Human agreement	92.2 %

(a) Confusion matrix (b) Accuracy

Table A1. **Humans vs. LLM-as-a-judge:** To validate that the LLM accurately identified objects in textual descriptions without hallucination, we provided human raters with a description and a single object. They then answered a Yes/No question: ‘*is the object mentioned in the text?*’ (a) Comparing LLM to human annotations. (b) Accuracy values for LLM. Note that even for such a simple task, the inter-human agreement is 92.2%.

sions from human raters. Out of this, we filtered over 100 impressions that were inconsistent with the majority vote of human annotators for the same question. Table A1(a) shows the confusion matrix between LLM and humans. Based on these values, we computed the true-positive rate (how accurately the LLM spotted objects in the descriptions) – 95.2%, the true-negative rate (the degree to which LLM avoided hallucinating objects) – 96.5%, and finally, the total accuracy – 95.7%. We also note that despite the simplicity of the task, human raters were not in full agreement; the user response agreement was 92.2% – on par with the LLM’s accuracy.

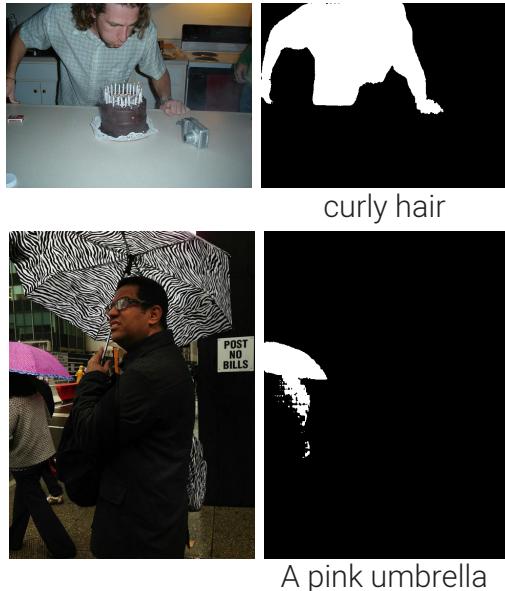


Figure A3. **Rejecting SAM masks:** An input image (left) and the corresponding SAM mask (right). The text used to prompt EVF SAM [39] is shown beneath each mask. We manually rejected these segmentation masks since they do not correspond well to the textual description or are of low quality.

C. Annotating Fine-Grained Details

Sec. 4.5 explored how the model retrieves fine-grained visual information from image tokens. For the purpose of this experiment, we defined a “fine detail” as a concrete object that was spotted by the baseline VLM but was omitted under $KO_{img \rightarrow gen}$ knockout setting. Sec. 4.3 showed that information conveying these objects is not being accumulated in the text query tokens, and the experiment in Sec. 4.5 was set to discover whether it comes by attending directly to image tokens. To answer this question, we annotated fine details in images from the same subset of COCO images (Sec. E). We considered all false-negative details extracted during the LLM-as-a-judge evaluation for our visual-to-output knockout experiment of Fig. A4(b) as candidate fine-grained visual details since the model was unable to describe them using the query text tokens alone. Note that these details are not restricted to any pre-defined set of categories but rather follow an “open vocabulary” setting where the details are defined based on analyzing differences in free-text image descriptions. Furthermore, since the details are derived from a specific knockout experiment, different VLM models induce different lists of candidate details. We further asked an LLM to associate each extracted detail with specific generated text tokens of the full description. Given the textual description of the details in the images, we used text-guided segment anything model [39] to create a binary mask localizing each detail in the image. Finally, we manually inspected the extracted details and their masks and discarded details for which the masks did not match the textual description, were poorly localized or were of low quality, see examples in Fig. A3.

After this manual selection, we were left with 231 annotated details in 68 images for InternVL2, and 115 details in 57 images for LLaVA-1.5. Each annotated detail comprises a segmentation mask, localizing it in *image* space, and a short textual description, localizing it in the generated *text*.

D. LLaVA-1.5 analysis

In this section we provide our analysis results on LLaVA-1.5-7B [24].

Attention Knockout Analysis We visualize in A5 the fraction of attention towards each token type: T_{img} , T_{txt} , T_{gen} . It depicts a non-uniform flow of information across layers, as shown for InternVL2 in Fig. 1.

We repeat our knockout experiments from Sec. 4 on LLaVA-1.5, and provide the results in A4 for both LLaVA-1.5 and InternVL2. We observe that all trends and observations from InternVL2 also occur in LLaVA-1.5: (a) the query tokens have an essential role as global image descriptors, (b) there is a special role for the mid-layers. Specifically, the mid-layers 4-20, which are only about 50% of the

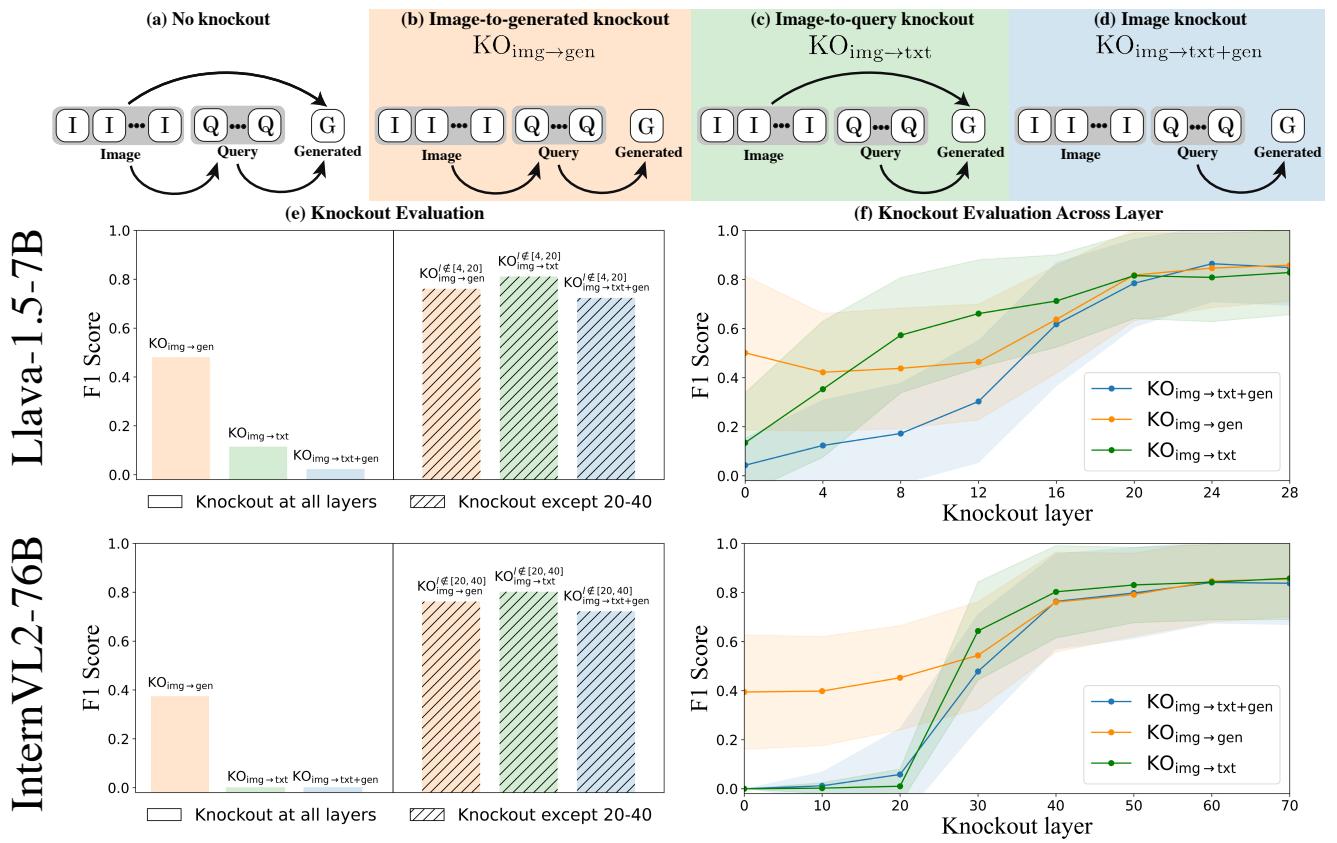


Figure A4. **Attention knockout on LLaVA-1.5 [24] and InternVL2 [6]:** (a) The VLM employs causal masking (Eq. 1), allowing generated and query tokens to gather information from image tokens, but not vice versa. We analyze three knockout configurations: (b) Image-to-generated $\text{KO}_{\text{img} \rightarrow \text{gen}}$: visual information flows to generated tokens only through query tokens, (c) Image-to-query $\text{KO}_{\text{img} \rightarrow \text{txt}}$: blocks query tokens from accessing image information, and (d) Image-to-others $\text{KO}_{\text{img} \rightarrow \text{txt+gen}}$: blocks image tokens from affecting all other tokens. (e) Evaluation of model responses (see Sec. 4.2) under each knockout configuration reveals that $\text{KO}_{\text{img} \rightarrow \text{gen}}$ achieves a 0.4 F1 score despite indirect image access, while $\text{KO}_{\text{img} \rightarrow \text{txt}}$ fails completely, demonstrating query tokens’ essential role as global image descriptors. (f) We expand previous experiments by knocking out attention, starting from layer l . Results highlight a consistent rise in F1 scores in the mid-layers, suggesting their critical role in visual information processing.

layers, are responsible for most part of the information flow between the image and text modalities. We note both models exhibit such redundancy (25% of the layers are sufficient in InternVL2, 50% for LLaVA-1.5), and we hypothesize the difference comes mainly from the fact that LLaVA-1.5 is much smaller in parameter size.

Top-K Image Tokens Importance Finally, in corresponds to Fig. 9 in the main paper, we turn to validate if the visual tokens also exhibit a redundancy, when evaluating the model’s performance while allowing only the top-k highest attended tokens to influence the generated tokens. Results are provided in Fig. A7, and indicates that for LLaVA-1.5 a redundancy exists as well. However, it saturates slower, and we hypothesize it is due to the fact that LLaVA-1.5 has much fewer visual tokens (256 vs 1600 on average for InternVL2), a difference which stems from the multi-resolution encoding strategy of InternVL2. Therefore, in

LLaVA-1.5, using 5% of the tokens is only 13 tokens, relative to 80 tokens in InternVL2.

Qualitative results for the different knockout settings, on the same images used in the main paper at Fig. 5, is provided in Fig. A8.

Fine-Grained Details Localized in Mid Layers Fig. A9 shows examples of the annotated details localized both in the image (segmentation mask) and in the generated text. The aggregated attention maps of the mid-layers (layers 16-24 for LLaVA-1.5), corresponding to the generated text tokens, show good localization of the details in the image.

Additionally, we report localization accuracy for the annotated details in Fig. A6. The trend is similar for both models – localization is done only in several middle layers.

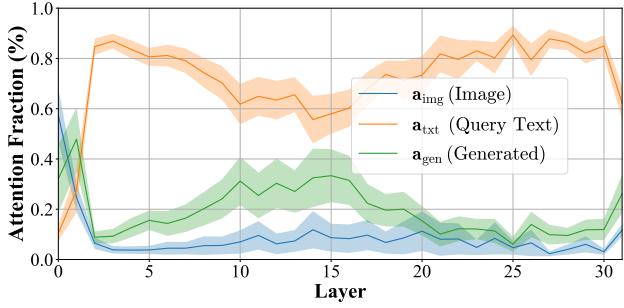
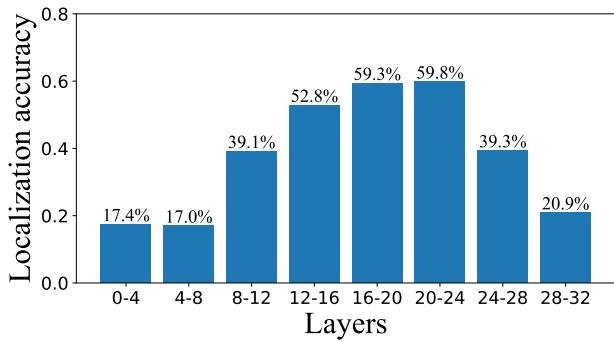
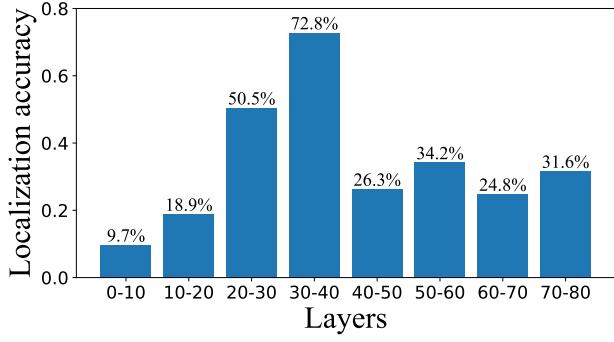


Figure A5. LLaVA-1.5 Fraction of attention to different token types: We measure the relative amount by which the generated tokens attend to: image tokens (blue), query text tokens (orange), and the previously generated tokens in the sequence (green). We report the distribution of relative attention for a set of 80 images, averaged across attention heads and generated tokens for LLaVA-1.5.



(a) LLaVA-1.5



(b) InternVL2

Figure A6. Object localization accuracy. We check if the attention of generated tokens associated with a specific object peak within one token distance from the pseudo ground truth object mask. We report the average accuracy across every four consecutive layers. (a) Results for LLaVA-1.5. (b) Results for InternVL2 (presented in Fig. 7 of the main paper and brought here for reference). The trend is similar for both models – localization is done only in several middle layers.

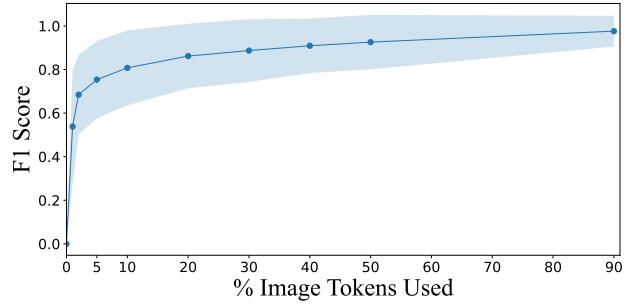


Figure A7. LLaVA-1.5 Image tokens redundancy: Similar to Fig. 9, we evaluate the model performance, while letting the generated tokens access to only the top-k image tokens with the highest attention values. It does show a similar trend, where a small percentage of the tokens are enough to provide a high F1 score above 0.8. Moreover, we note that as LLaVA-1.5 accepts only single-resolution patch, it has far less tokens, and 5% in this case maps to only 13 tokens.

E. Images used for evaluation

All evaluations in the paper were conducted using a subset of 81 images from the COCO [23] dataset. Fig. A12 shows the photos selected and their IDs. The images depict complex scenarios of various indoor and outdoor scenes with many fine details. To evaluate Image Reprompting (Sec. 5), we used the MME dataset [11].



The image features a group of pigeons gathered on a paved surface, with one of them sitting on top of a piece of bread. The pigeon is positioned in the center of the scene, while the other pigeons are scattered around it. Some of the pigeons are standing close to the pigeon with the bread, while others are further away. The pigeons appear to be enjoying their time together, possibly sharing the bread as a snack.



(a) Baseline model: no knockout

The image features a man in a military uniform, possibly a Marine, standing outside and looking at his cell phone. He is holding the phone in his hand, possibly checking messages or browsing the internet. There are several other people in the scene, some of whom are also wearing military uniforms. One person is standing close to the man with the cell phone, while others are scattered around the area. A bottle can be seen on the ground, possibly belonging to one of the individuals in the scene.



The image features a black cat sitting inside a white bathroom sink. The sink is located next to a toilet, and there are several bottles scattered around the sink area. Some of these bottles are placed on the counter, while others are located near the sink. The scene appears to be a cozy and unusual resting spot for the cat.

(b) Knockout image-to-gen, $KO_{img \rightarrow gen}^{(\notin [20,40])}$, except layers 20-40

The image features a man in military fatigues, possibly a Marine, wearing a hat and holding a cell phone. He is standing outside, possibly in a park, and appears to be looking at his phone. There are several other people in the background, some of whom are also wearing hats. A few of them are standing close to each other, while others are more spread out. Additionally, there is a bottle visible in the scene, possibly belonging to one of the people present.

The image features a group of pigeons gathered on a street, with one of them sitting on top of a piece of bread. The pigeon is positioned in the center of the scene, surrounded by several other pigeons. Some of the pigeons are standing, while others are sitting or walking around. The street appears to be a mix of pavement and dirt, creating a casual and urban setting for the pigeons.

True Positive (TP): Pigeon, Break
False Negative (FN): N/A
False Positive (FP): N/A
Precision=2/2 Recall= 2/2

F1=1.0

TP: Man, People, Military uniform/fatigues, Cell phone, Bottle
FN: N/A **FP:** Hat
Precision=5/6 Recall= 5/5 **F1=0.90**

The image features a black cat sitting inside a bathroom sink, seemingly enjoying the cozy space. The sink is located next to a counter, and there are several bottles scattered around the area. Some of these bottles are placed on the counter, while others are located near the sink. The scene gives off a relaxed and comfortable atmosphere, as the cat seems to be at ease in the bathroom environment.



The image features a group of pigeons gathered in a courtyard. There are at least nine pigeons visible in the scene, with some standing closer to the foreground and others further back. The pigeons are scattered throughout the courtyard, with some standing near the center and others closer to the edges. The scene captures a typical moment of urban wildlife interacting with their surroundings.

True Positive (TP): Pigeon
False Negative (FN): Bread
False Positive (FP): N/A
Precision=1/1 Recall= 1/2

F1=0.67

(c) Knockout image-to-gen, $KO_{img \rightarrow gen}$, for all layers

The image features a man in a military uniform, possibly a soldier, standing in front of a building. He is wearing a backpack and appears to be looking at something in the distance. The soldier is positioned towards the left side of the image. There are two other people in the scene, one located in the middle and the other on the right side. The middle person is closer to the soldier, while the person on the right is further away. The scene seems to be taking place outdoors, with the building in the background.

TP: Man, People, Military Uniform
FN: Cell phone, Bottle
FP: Backpack, Building
Precision=3/5 Recall= 3/5 **F1=0.60**

The image features a black and white cat sitting on a chair in a room. The cat appears to be looking at something, possibly a camera or another object in the room. The chair is positioned in the center of the scene, and the cat is comfortably occupying it.



The image features a person standing in front of a building, possibly a store or a shop. The person is wearing a tie, which suggests a professional or formal attire. The building appears to be a storefront, with the person standing in front of it, possibly waiting for customers or preparing to enter the establishment.

Precision=0 Recall= 0

F1=0.0

(d) Knockout image-to-text, $KO_{img \rightarrow txt}$, for all layers

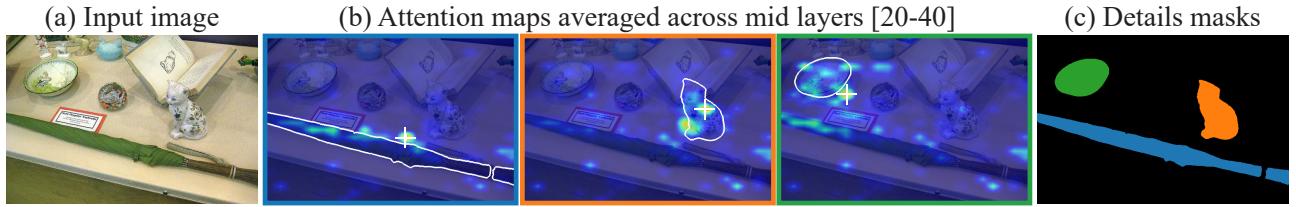
The image features a person standing in front of a computer screen, likely working or browsing the internet. The person is wearing a tie, which suggests a professional or formal setting. The computer screen takes up a significant portion of the image, indicating its importance in the scene. The person appears to be focused on the screen, possibly engaging in tasks such as writing an email.

Precision=0 Recall= 0 **F1=0.0**

The image features a black and white cat sitting on a table. The cat appears to be looking at something, possibly a camera or another object of interest. The table is positioned in the center of the scene, with the cat occupying a significant portion of the image. The focus of the image is on the cat and its surroundings, creating a sense of curiosity and engagement with the viewer.



Figure A8. Qualitative results for knockout experiments on LLaVA-1.5: We use our LLM-as-a-judge protocol, \ddagger , to compare the baseline VLM description of images (a) to descriptions generated under various attention knockouts. (b) Allowing generated tokens to attend to image tokens only in mid-layers 20-40, $KO_{img \rightarrow gen}^{(\notin [20,40])}$, does not degrade the description significantly – F1 scores are close to 1.0. (c) Blocking attention between generated and image tokens for all layers, $KO_{img \rightarrow gen}$, results in loss of fine details, e.g., the bagel, smartphone or the toothpaste, and hallucinations, e.g., a black cap for the officer. Consequently, F1 scores are significantly lower – around 0.45. (d) When blocking attention between query text and image tokens for all layers, $KO_{img \rightarrow txt}$, the VLM is no longer able to describe the image – F1=0. We note that LLM evaluation can be noisy, leading to slight inconsistencies in the identified objects across different comparisons. For instance, in the rightmost examples, (b) and (c) show variations in the number of identified objects in the baseline (4 and 5).



The image features a wooden table with a variety of items on it. There is a book placed on the table, along with **a green umbrella**, **a cat figurine**, and **a bowl**. The cat figurine is positioned near the book, while the bowl is located closer to the edge of the table. The green umbrella is placed in the middle of the table, creating a visually interesting arrangement of objects.



The image features an older woman standing in a kitchen, preparing food on a dining table. She is focused on making **bread** and rolls, with a variety of dough and ingredients spread out on the table. The kitchen is well - equipped with a refrigerator, an oven, and a sink. There are **several chairs** placed around the dining table, and **a couch** can be seen in the background. Additionally, there are **multiple cups** and a bowl on the table, possibly containing ingredients or be ver ages .



The image captures a large airplane flying low over **a field**, with **a car** and **a group of people** visible below. The airplane is positioned towards the left side of the scene, while the car is located on the right side, closer to the bottom. The people are scattered around the field, with some standing closer to the car and others further away. The scene appears to be a mix of an airplane taking off or landing and people observing the event from the ground.



The image features a woman sitting on **a motor scooter**, which is loaded with **a basket** full of **fresh vegetables**. She appears to be preparing to ride the scooter with her produce. There are several other people in the scene, some of whom are standing or walking around, while others are sitting on a bench. In addition to the motor scooter, there are two cars visible in the background, one on the left side and another on the right side of the image. A back pack can also be seen placed on the ground near the center of the scene.

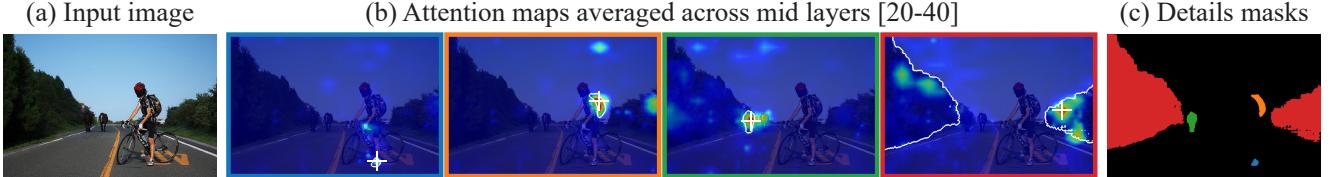


The image features a room with **a large window**, show casing a beautiful beach scene. **A banana** is placed in a yellow chair, positioned in front of the window, as if it is enjoying the view. The chair is placed next to a small table, and there are two other chairs in the room, one on the left side and another on the right side. In addition to the chairs, there are **two umbrellas** in the room, one located near the center and the other towards the right side. A remote control can be seen on the table, and a book is placed on the left side of the room. The overall scene creates a cozy and relaxing atmosphere, as if the banana is taking a break from its daily routine to enjoy the beach view.



The image depicts a woman walking down a street in a residential area. She is carrying a hand bag and appears to be crossing the street at **a stop sign**. The stop sign is located near the center of the scene, and the woman is walking towards it. There are several cars parked along the street, with one car on the left side of the scene, another car further down the street, and a third car on the right side. Additionally, there is **a fire hydrant** situated near the center of the scene, and **a tree** can be seen in the background, adding to the residential atmosphere.

Figure A9. Attending to objects: Results for LLaVA-1.5. (a) Input image. (b) Average attention maps of the generated tokens associated with each object (marked in color in the generated text). (c) Pseudo ground truth object masks, generated using SAM [20, 39]. The peak of attention, marked by a white cross, aligns with the location of the object in the image, not as well as for InternVL2.



The image depicts a cyclist riding a road bike on a paved road. The cyclist is wearing a red helmet, black and white cycling jersey, black shorts, and **white cycling shoes**. They are also carrying a **black backpack**. The road is marked with a double yellow line down the center and a white line along the edges. On the left side of the road, there are **two horses** walking in the same direction as the cyclist. The surrounding area is green with trees and **bushes** on both sides of the road. The sky is clear and blue, indicating good weather conditions.



The image shows a display table with various items arranged on it. The items include:

1. A green umbrella with a wooden handle,
2. **A small figurine of a cat with a floral pattern**
3. **A decorative bowl with a design on it**
4. **A small snow globe with a colorful scene inside**
5. A small blue bird figurine
6. A small white figurine of a person holding a white animal (possibly a sheep or a lamb)
7. **An open book with an illustration of a cat** on the left page and text on the right page
8. **A red and white informational card** placed in front of the book,
9. A small green plant in a pot,
10. A small decorative item that appears to be a blue bird or animal figurine.

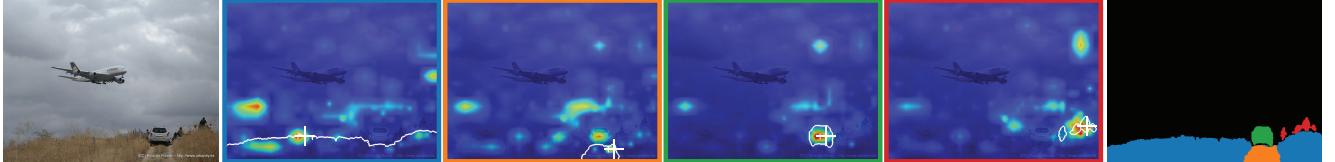
The table seems to be part of an exhibit or display,



The image shows a well-lit, kitchen with modern appliances and wooden flooring. The kitchen features **a large island** in the center with a wooden countertop, and **three high stools with black frames and light gray cushions**. On the island, there is **a vase with colorful flowers**, and a bowl of fruit. The kitchen has light wooden cabinets both above and below the countertops. The upper cabinets are open, and display various dishes and kitchen items. There is a **stainless steel refrigerator** on the left side of the image, with **some papers** and magnets attached to it. The back wall has a window with blue shutters, and **a door with a window**, that appears to lead outside. There are various kitchen appliances and utensils on the countertops, including a coffee maker, and a toaster. The overall atmosphere of the kitchen is clean, and organized.



The image shows a small dog, **on a leash**, being walked by a person wearing **black high-heeled shoes**. The dog is wearing **a yellow bandana** and **a white shirt with a colorful design on the back**. The dog appears to be sniffing or licking the ground. The setting seems to be a sidewalk or paved area.



The image depicts a large Lufthansa airplane in mid-flight, during what appears to be a landing approach. The aircraft is flying low over **a field** with tall, grass, and a few scattered trees. Below the airplane, is **a dirt road** with **a white SUV** driving on it. On the right side of the image, are **several people standing and watching** the airplane. The sky is overcast with thick, gray clouds, covering the entire sky, background.

Figure A10. Attending to objects: Completing results for InternVL2 shown in Fig. 6 of the main paper. (a) Input image. (b) Average attention maps of the generated tokens associated with each object (marked in color in the generated text). (c) Pseudo ground truth object masks, generated using SAM [20, 39]. The peak of attention, marked by a white cross, aligns well with the location of the object in the image.

Perception (Coarse-Grained Tasks)

Existence

[Y] Is there a elephant in this image?	[Y] Is there a refrigerator in this image?
[N] Is there a hair drier in this image?	[N] Is there a donut in this image?

Count

[Y] Is there a total of two person appear in the image?	[Y] Are there two pieces of pizza in this image?
[N] Is there only one person appear in the image?	[N] Is there only one piece of pizza in this image?

Position

[Y] Is the motorcycle on the right side of the bus?	[Y] Is the baby on the right of the dog in the image?
[N] Is the motorcycle on the left side of the bus.	[N] Is the baby on the left of the dog in the image?

Color

[Y] Is there a red coat in the image?	[Y] Is there a red couch in the image?
[N] Is there a yellow coat in the image?	[N] Is there a black couch in the image?

Perception (Fine-Grained Tasks)

Poster

[Y] Is this movie directed by francis ford coppola ?	[Y] Is this movie titled twilight (2008) ?
[N] Is this movie directed by franklin j. schaffner ?	[N] Is this movie titled the horse whisperer (1998) ?

Celebrity

[Y] Is the actor inside the red box called Audrey Hepburn ?	[Y] Is the actor inside the red box named Jim Carrey ?
[N] Is the actor inside the red box called Chris April ?	[N] Is the actor inside the red box named Jari Kinnunen ?

Scene

[Y] Does this image describe a place of moot water ?	[Y] Is this picture captured in a place of galley ?
[N] Does this image describe a place of marsh ?	[N] Is this picture captured in a place of physics laboratory ?

Landmark

[Y] Is this an image of Beijing Guozijian ?	[Y] Is this a picture of Church of Saint Giles in Prague ?
[N] Is this an image of Klinikkirche (Pfafferode) ?	[N] Is this a picture of Pfarrkirche St. Martin an der Raab ?

Artwork

[Y] Does this artwork belong to the type of still-life ?	[Y] Is this artwork displayed in musée du louvre ?
[N] Does this artwork belong to the type of mythological ?	[N] Is this artwork displayed in galleria nazionale d'arte moderna e contemporanea ?

Perception (OCR Task)

OCR

[Y] Is the phone number in the picture " 0131 555 6363 "?	[Y] Is the word in the logo " high time coffee shop "?
[N] Is the phone number in the picture " 0137 556 6363 "?	[N] Is the word in the logo " high tite cofeee shop "?

Figure A11. **MME perception tasks:** Illustration of the different tasks of the MME benchmark, taken from [11] (cf Fig. 1). MME contains ten perception tasks. Each image is associated with two questions whose answers are marked yes [Y] or no [N], respectively. The instruction consists of a question followed by “Please answer yes or no”. Results over all subsets are provided in Tab. 1

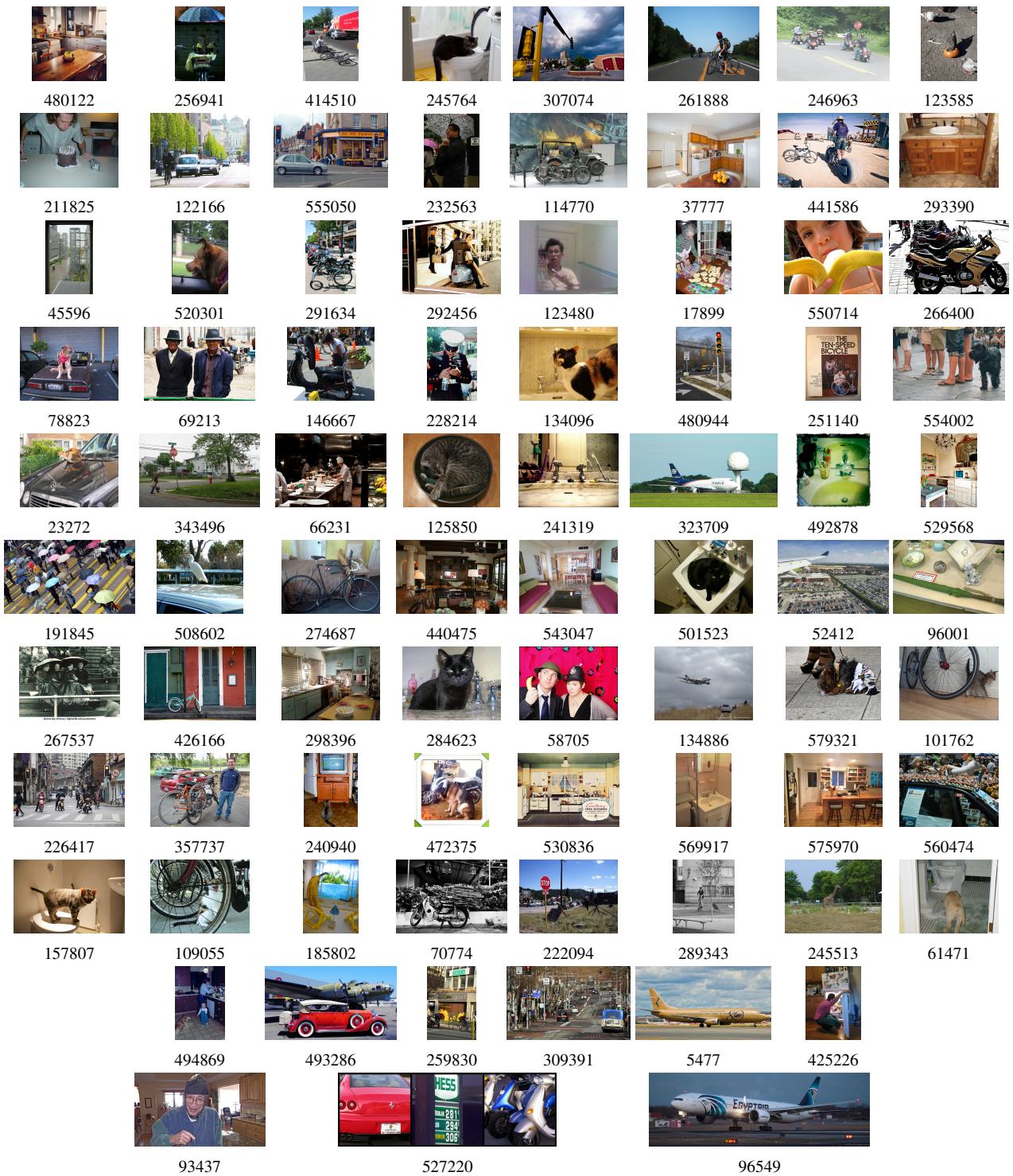


Figure A12. **Images selected for VLM inspection.** We used the following images from COCO [23] depicting complex and varied scenes. Beneath each image appears its COCO ID.