

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

Supplementary Material

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

*Indicates equal contribution.

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

Contents

A VLMs in Our Analysis	1
B LLM-as-a-judge	1
C Annotating Fine-Grained Details	3
D LLaVA-1.5 analysis	4
E InternVL2 Additional Results	6
F. Images used for evaluation	6

the main paper. Specifically, given original and modified VLM’s responses (e.g., before and after knock-out), 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.

A. VLMs in Our Analysis

We use InternVL2-76B [1] – a powerful open-source Visual Language Model built upon Llama3-70B LLM [3] and a 6B ViT encoder [2]. InternVL2 demonstrates highly competitive performance, surpassing other open-source VLMs, including LLaVA models, and achieving results comparable to closed-source models across multiple benchmarks [1, 8]. We further validate our results using LLaVA-1.5-7B [7], 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 process high resolution images by splitting them to several high-res non-overlapping patches, alongside a one low-res patch of the resized image. This should enable the model to extract finer details. However, LLaVA-1.5 simply resize 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

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, 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.

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 impressions 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.

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 to-

LLM	Humans		True-positive rate	95.2 %
	Yes	No	True-negative rate	96.5 %
			Total accuracy	95.7 %
	No		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%.

kens, 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. F). 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

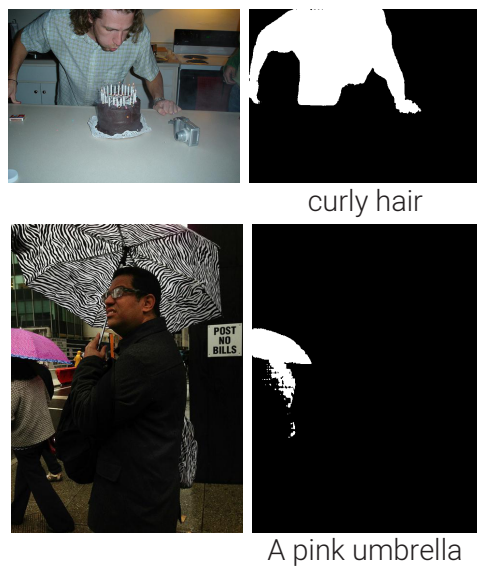


Figure A3. **Rejecting SAM masks:** An input image (left) and the corresponding SAM mask (right). The text used to prompt EVF SAM [9] is shown beneath each mask. We manually rejected these segmentation masks since they do not correspond well to the textual description or are of low quality.

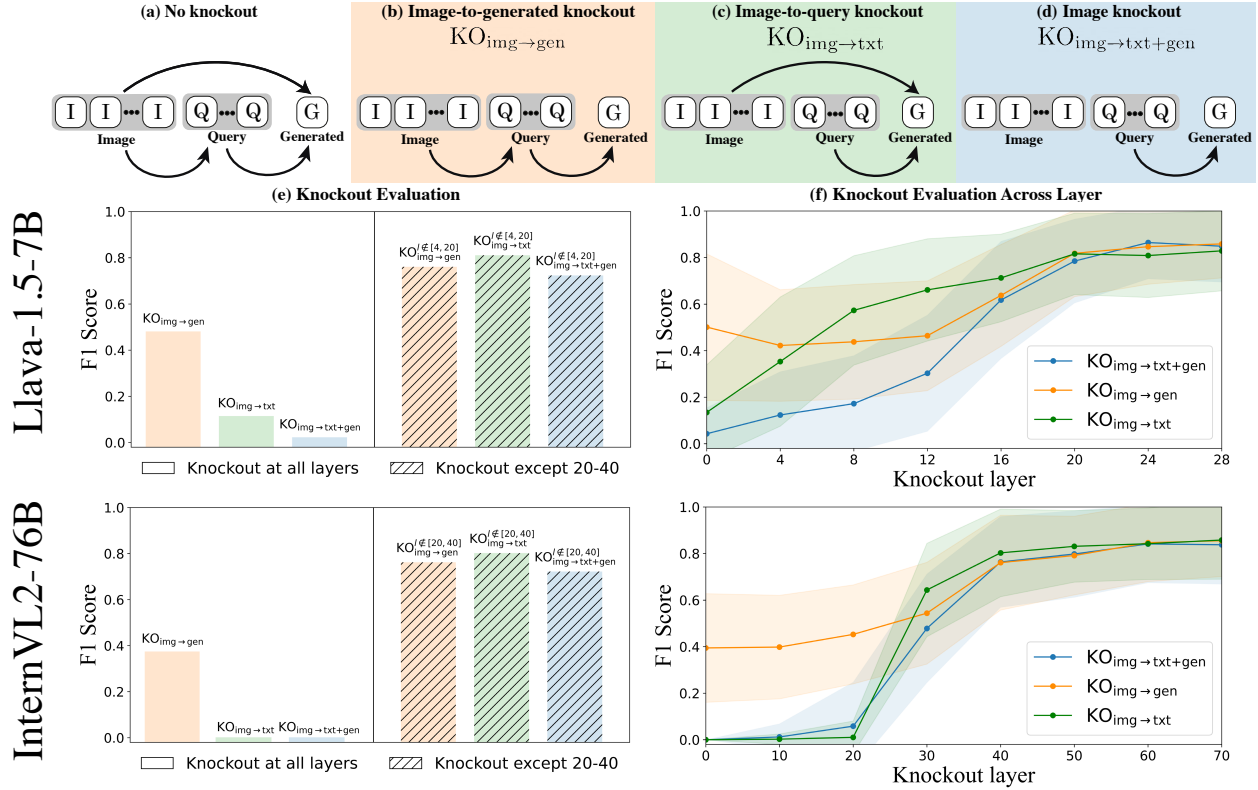


Figure A4. **Attention knockout on LLaVA-1.5b [7] and InternVL2 [1]:** (a) The VLM employs causal masking (Eq.1), allowing generated and query tokens to gather information from image tokens, but not vice versa. We analyze three knockout configurations: (b) Image-to-generated $KO_{img \rightarrow gen}$: visual information flows to generated tokens only through query tokens, (c) Image-to-query $KO_{img \rightarrow txt}$: blocks query tokens from accessing image information, and (d) Image-to-others $KO_{img \rightarrow txt+gen}$: blocks image tokens from affecting all other tokens. (e) Evaluation of model responses (see Sec. 4.2) under each knockout configuration reveals that $KO_{img \rightarrow gen}$ achieves a 0.4 F1 score despite indirect image access, while $KO_{img \rightarrow txt}$ 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.

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 [9] 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 [7].

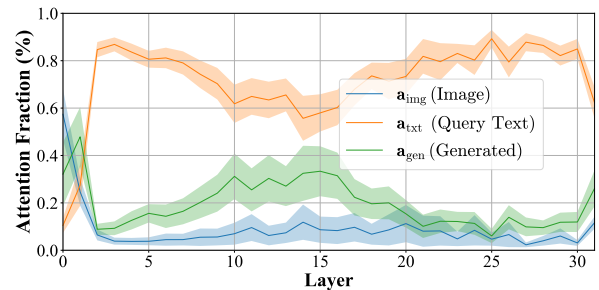
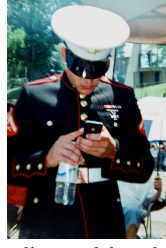


Figure A5. **LLaVA-1.5 Fraction of attention to different token types:** We measure the relative amount by which the generated tokens attend to: image tokens (blue), query text tokens (orange), and the previously generated tokens in the sequence (green). We report the distribution of relative attention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

Attention Knockout Analysis We visualize in A5 the fraction of attention towards each token type: T_{img} ,



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, seemingly enjoying the cozy space. 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 countertop, 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, $KO_{img \rightarrow gen}^{[20,40]}$, except layers 20-40

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**

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.

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.

TP: Cat, Sink, Bottles, Counter
FN: Toilet
FP: N/A
 Precision=4/4 Recall= 4/5 **F1=0.89**

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

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**

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.

TP: Cat
FN: Sink, Toilet, Bottles
FP: Chair
 Precision=1/2 Recall= 1/4 **F1=0.33**

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

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**

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.

Precision=1/2 Recall= 1/5 **F1=0.28**

Figure A6. Qualitative results for knockout experiments on LLaVA-1.5: We use our LLM-as-a-judge protocol, , 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, $KO_{img \rightarrow gen}^{[20,40]}$, 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 (6 and 7).

T_{txt}, T_{gen} . It depicts a non-uniform flow of information across layers, as shown for InternVL2 in Fig. 1 in the main paper.

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

Specifically, the mid-layers 4-20, which are only about 50% of the 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. A8, 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 less 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. A6.

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. A7. The trend is similar for both models – localization is done only in several middle layers.

E. InternVL2 Additional Results

This section provides additional results for InternVL2-76B [2].

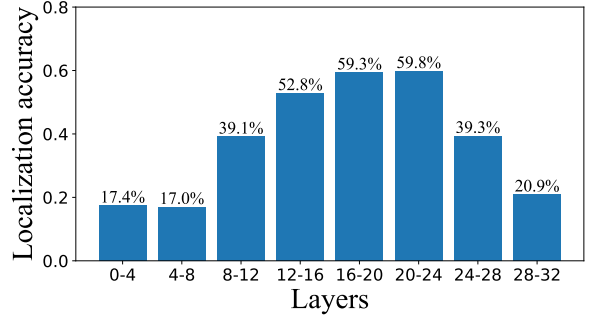
Fine-Grained Details Localized in Mid Layers

Fig. 6 in the main paper and Fig. A10 show additional localization results for InternVL2. The figures include the full generated text and highlight the specific text tokens *automatically* associated with each detail.

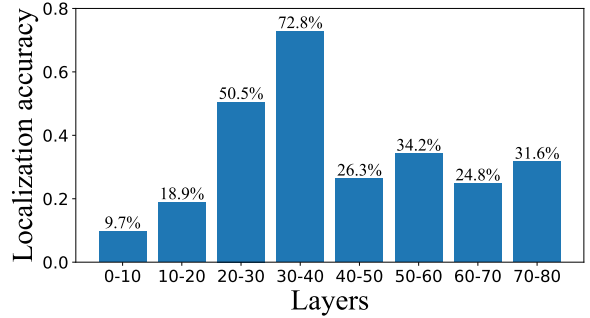
Image Re-prompting We extend our evaluation for compressed context of InternVL2 on MME [4] to all 10 perception sub-tasks illustrated in A11, and provide results in Tab. A2. We observe that while using only 6% of all tokens in the sequence, the compressed context achieves almost the same performance as the *Naive* baseline (i.e., prompting the model for each question-image pair independently), and surpassing the *Describe-to-LLM* baseline.

F. Images used for evaluation

All evaluations in the paper were conducted using a subset of 81 images from the COCO [6] dataset. Fig. A12



(a) LLaVA-1.5



(b) InternVL2

Figure A7. **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.

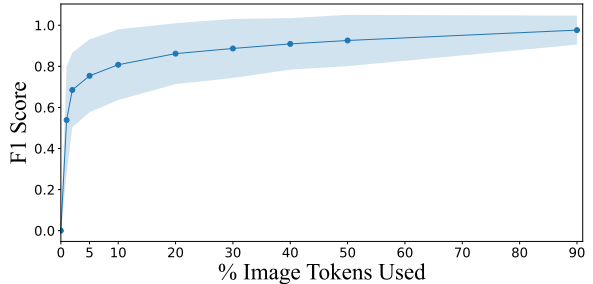
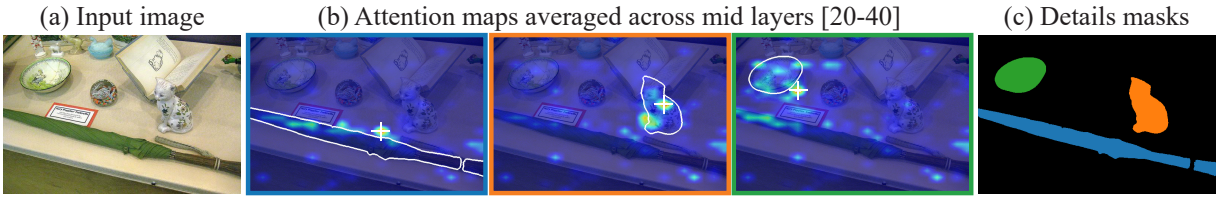
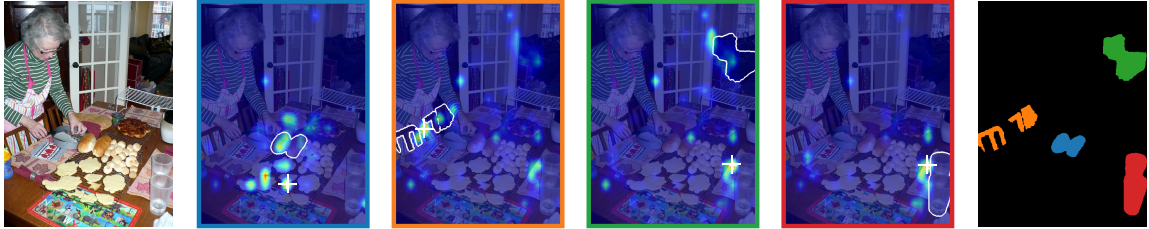


Figure A8. **LLaVA’s Image tokens redundancy:** Similar to Fig. 9, we evaluate the model performance, while letting the generated tokens access to only the top-k image tokens with the highest attention values. It does show a similar trend, where a small percentage of the tokens are enough to provide a high F1 score above 0.8. Moreover, we note that as LLaVA accepts only single-resolution patch, it has far less tokens, and 5% in this case maps to only 13 tokens.

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 Re-prompting (Sec. 5), we used the MME dataset [4].



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 .



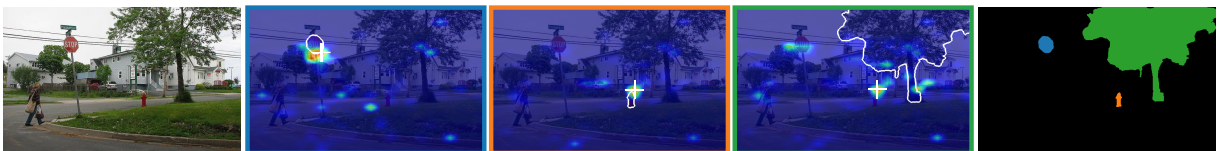
The image captures a large airplane flying low over a field, with a car and a group of people visible below. The airplane is positioned towards the left side of the scene, while the car is located on the right side, closer to the bottom. The people are scattered around the field, with some standing closer 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 [5, 9]. 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.

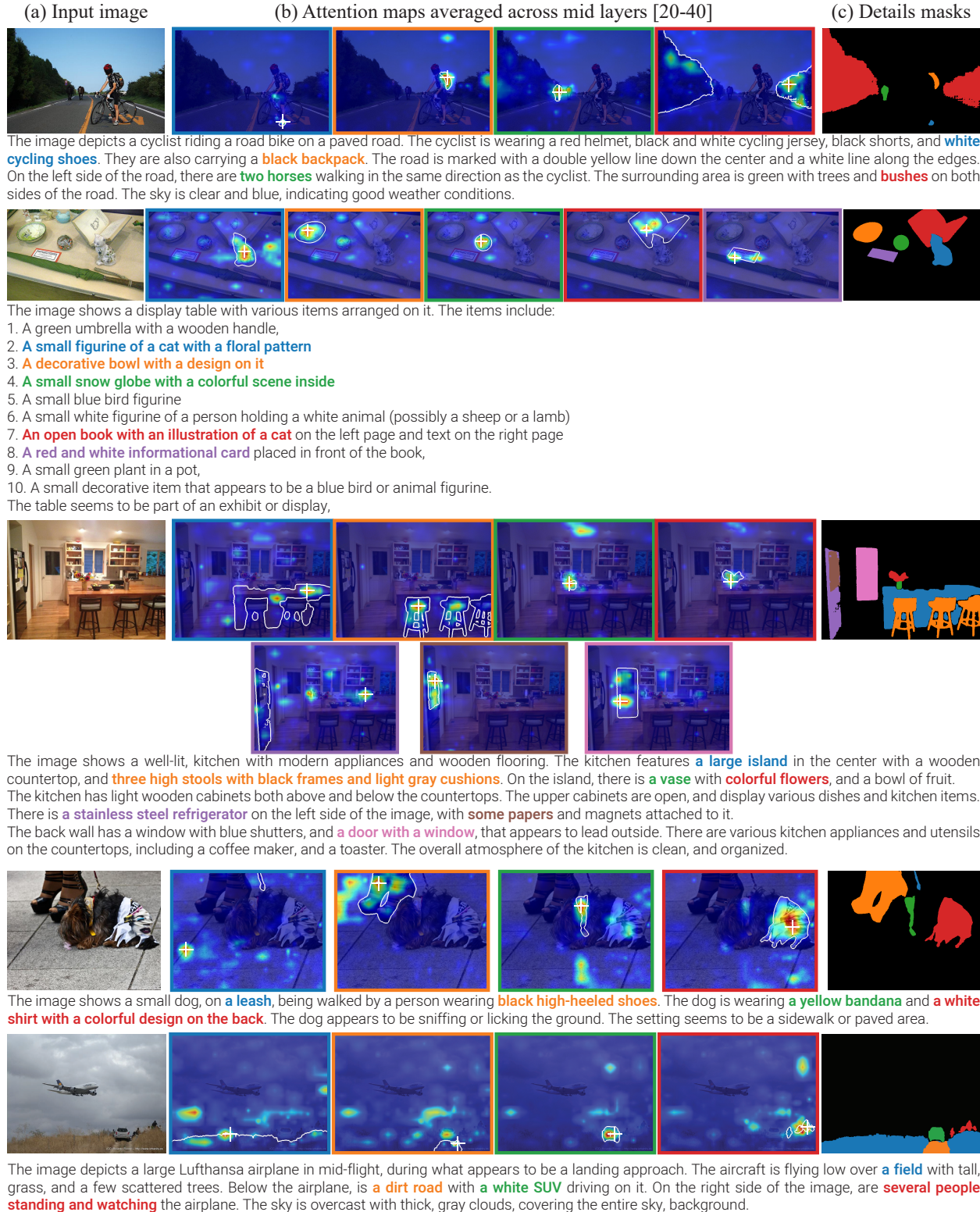


Figure A10. Attending to objects: Completing results for InternVL2 shown in Fig. 6 of the main paper. (a) Input image. (b) Average attention maps of the generated tokens associated with each object (marked in color in the generated text). (c) Pseudo ground truth object masks, generated using SAM [5, 9]. The peak of attention, marked by a white cross, aligns well with the location of the object in the image.



Figure A11. **MME perception tasks:** Illustration of the different tasks of the MME benchmark, taken from [4] (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. A2

		Existence		Count		Position		Color		OCR		Poster	
		ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+
Compressed Context	Naive (InternVL2)	98.33	96.77	81.67	63.33	80.00	60.00	86.67	76.67	67.50	35.00	90.13	84.35
	Describe-to-LLM	90.00	80.00	75.00	73.33	66.67	46.67	86.67	80.00	77.50	55.00	86.05	80.27
	Query + K=5%	91.66	83.33	85.00	70.00	68.33	40.00	80.00	60.00	77.50	55.00	87.55	78.23
	Query + K=2%	85.00	76.67	78.33	60.00	68.33	40.00	70.00	40.00	72.50	45.00	82.39	65.49
	Query	56.67	13.33	46.67	13.33	53.33	16.67	46.67	3.33	55.00	15.00	71.08	48.29
	K=2%	65.00	30.00	56.67	30.00	56.67	30.00	50.00	10.00	52.50	10.00	64.28	33.33
		Celebrity		Artwork		Scene		Landmark		Average		Reprompt	
		ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	ACC	ACC+	#Tokens	
Compressed Context	Naive (InternVL2-76B)	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	
	Query + K=5%	79.41	58.83	84.67	71.85	83.00	67.50	78.94	60.52	81.46	64.52	201	
	Query + K=2%	77.94	56.47	83.50	69.50	80.25	61.50	71.92	49.12	77.16	55.94	151	
	Query	70.00	41.76	72.50	49.50	73.50	50.00	64.91	33.33	61.03	28.45	60	
	K=2%	52.05	10.00	78.00	61.00	80.50	62.00	68.42	42.10	62.40	31.84	91	

Table A2. **Evaluation on MME (InternVL2-76B):** The results cover 10 Perception tasks of the MME benchmark [4]. Metrics include accuracy (ACC), ACC+ (percentage of images where all questions are correct), and the number of tokens used for re-prompting. The sum of ACC and ACC+ is the total score reported in the main paper. 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-prompting an image (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.

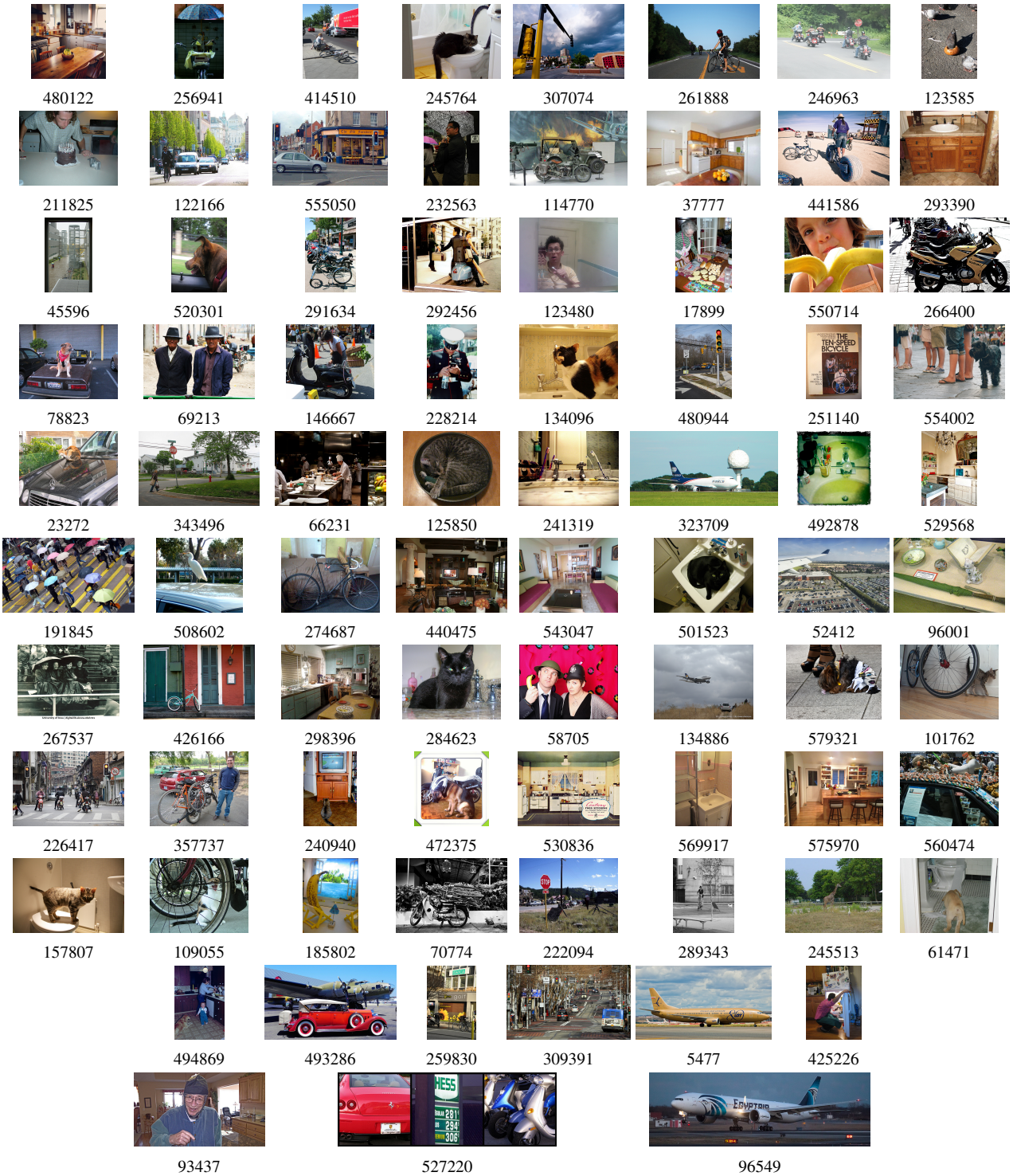


Figure A12. **Images selected for VLM inspection.** We used the following images from COCO [6] depicting complex and varied scenes. Beneath each image appears its COCO ID.

References

- [1] 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](#), [4](#)
- [2] Zhe Chen, Jiannan Wu, Wenhai 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](#), [6](#)
- [3] 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. [1](#)
- [4] 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. [6](#), [9](#)
- [5] 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. [7](#), [8](#)
- [6] 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. [6](#), [10](#)
- [7] 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](#), [4](#)
- [8] Chufan Shi, Cheng Yang, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, Gongye Liu, Xiaomei Nie, Deng Cai, and Yujiu Yang. Chartmimic: Evaluating lmm’s cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024. [1](#)
- [9] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xingang Wang. EVF-SAM: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. [3](#), [4](#), [7](#), [8](#)