Supplementary material for: Explaining Multi-modal Large Language Models by Analyzing their Vision Perception

Loris Giulivi Politecnico di Milano

Giacomo Boracchi

1 Training W

As discussed in Section 3.1, we train an alignment layer W between OWL's vision embedding and LLaVa's language model. The objective of the alignment layer is to transform the output of the OWL-ViT vision encoder *token-wise* to render it understandable by LLaVa. To achieve this result, we select as loss function the L_2 distance between the output transformed tokens $W(\mathcal{E}^{OWL}(\mathbf{x}))$ and the output from LLaVa's original vision encoder $\mathcal{E}^{I}(\mathbf{x})$:

$$\mathcal{L}(W,\mathbf{x}) = |\mathcal{E}^{I}(\mathbf{x}) - W(\mathcal{E}^{OWL}(\mathbf{x}))|_2.$$

The proposed procedure enables self-supervised training, where the input is composed by images $\mathbf{x}$ and the target is the output of the pre-trained LLaVa vision encoder $\mathcal{E}^{I}$, which is given. To ensure that W can correctly align a wide variety of images, we train on the large scale Open Images [1] dataset, including approximately 9 million images. Due to computational constraints, we train for a single epoch. Nonetheless, the large size of the dataset, combined with the relatively straightforward task and W's model architecture, still ensures that good alignment performance can be achieved, as demonstrated in Table 1.

The chosen architecture for W is a 2-hidden layer MLP with hidden dimensionality 8192. The input and output dimensionalities are 768 and 4096, conforming with OWL-ViT [3] and LLaVa [2]'s token dimensionalities, respectively. The alignment is applied to each vision token independently, such that given an input $B \times S \times 768$, we obtain output $B \times S \times 4096$, where the sequence length S is typically 576, which is the number of output vision tokens from OWL-ViT and equal to the sequence length required by the pre-trained LLaVa model. Since the alignment layer is applied separately for each token, for the purposes of loss computation, each token is treated as an independent sample.

In our work, due to limitations in computational resources, we relied on training W to construct the joint model $\mathcal{J}$. However, we hypothesize that it would be possible to fine-tune the LLaVa model directly to process the OWL-ViT vision encoder output without an alignment layer. This would potentially lead to better performance, since the alignment inherently leads to information loss, as shown in the main paper, Table 1. Additionally, we experimented with alternative architectures for the alignment layers, including the original LLaVa's alignment layer architecture and with a more complex transformer applied to the entire sequence[1], but with poor results, as shown in Table 1. Model weights are publicly available at: https://github.com/loris2222/ExplainingMLLMs.

Another potential way to construct $\mathcal{J}$ without requiring an alignment layer between the vision encoder and the language model, thus retaining the MLLM's full performance, would be to align LLaVa's vision encoder $\mathcal{E}^I(\mathbf{x})$ to OWL-ViT's classification and detection heads $\mathcal{H}$. This is effectively the opposite of the procedure discussed in this work. In this way, the MLLM pipeline could operate without changes, and alignment would only affect the performance of $\mathcal{H}$. We have experimented with this setup, but we were not able to train a reasonably functional alignment layer. This poor performance can be attributed to the nature of LLaVa's CLIP vision encoder, which is not suitable for object detection. Indeed, as stated from the original paper [3], OWL-ViT's vision encoder is an extensively modified version of CLIP, with a far greater input image size (768 vs 336) and peculiar training and architectural modifications, such as CLS token merging, which enable it to solve the detection task.

1.1 Evaluation details: As discussed in the paper in Section 3.1, we follow the gpt-as-a-judge paradigm [2] to evaluate alignment layers. We first generate captions from all models to be evaluted for the same 100 random COCO images, using prompt:

> `USER:` "*What's the content of the image?*"

We then prompt a strong judge (OpenAI's GPT4-vision) to score the provided captions:

> `SYSTEM:` "*Please act as an impartial judge and evaluate the quality of the caption provided by an AI assistant for the provided image. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, and level of detail of the caption. Be as objective as possible.*"
> `USER:` "*From 1 to 10, score this caption for the image: <caption>. Only answer with a number from 1 to 10.*"

We then average the 100 GPT4-provided scores for the captions generated by each model to be evaluated, and obtain the results of Table 1.

Table 1: Evaluation of alignment layer using gpt-as-a-judge [2]. Reporting the average GPT-4 score for 100 COCO captions provided by different models.

Model:	COCO GT	LLaVa	Ours (deep MLP)	Ours (LLaVa MLP)	Ours (transformer)
Avg. score:	6.9	**7.0**	6.1	5.6	3.2

2 Model output examples

In the following figures, we present examples of model outputs to multi-modal inputs $(\mathbf{x}, \mathbf{p})$, demonstrating that the alignment layer enables the LLM to understand vision information provided by OWL-ViT [3]. Prompts are taken from LLaVA-Instruct-150K [2]. We highlight in red discrepancies between GT and original LLaVa output with $\mathcal{J}$ due to imperfect alignments. If there are any, we highlight the detection for the relevant objects. From Figures 1- 4, it is clear how the detection output can help explain the LLM output.

[1]In this case, the loss cannot be computed by considering tokens as independent samples, since a transformer-based W can only be applied to the entire sequence. In this peculiar case, the loss can however still be computed as the average L_2 between the original and transformed vision tokens.

GT:

USER: What are the colors of the bus in the image?
ASSISTANT: The bus in the image is white and red.
USER: What feature can be seen on the back of the bus?
ASSISTANT: The back of the bus features an advertisement.
USER: Is the bus driving down the street or pulled off to the side?
ASSISTANT: The bus is driving down the street, which is crowded with people and other vehicles.

Original LLaVa:

USER: What are the colors of the bus in the image?
ASSISTANT: The bus in the image is red, white, and yellow.
USER: What feature can be seen on the back of the bus?
ASSISTANT: There is a large advertisement on the back of the bus.
USER: Is the bus driving down the street or pulled off to the side?
ASSISTANT: The bus is driving down the street.

$\mathcal{J}$:

USER: What are the colors of the bus in the image?
ASSISTANT: The bus in the image is red and white.
USER: What feature can be seen on the back of the bus?
ASSISTANT: The back of the bus has a large red heart.
USER: Is the bus driving down the street or pulled off to the side?
ASSISTANT: The bus is driving down the street.

Figure 1: Model $\mathcal{J}$ answers regarding a red heart instead of the advertisement on the back of the bus. As the detection output indicates, the graphic on the advertisement may resemble a red heart, which is identified by both $\mathbf{O}^{MLLM}$ and $\mathbf{O}^{OWL}$.

GT:

USER: Where is the cat positioned in the image?
ASSISTANT: The cat is positioned on top of the back of the couch in the living room.
USER: What is the cat doing in the image?
ASSISTANT: The cat is coming out from some curtains onto the couch and is sitting or standing on top of it.
USER: What is the color of the curtains near the cat?
ASSISTANT: The color of the curtains near the cat is red.
USER: Besides the cat, what else can be seen on the couch?
ASSISTANT: There are lots of pillows on the couch besides the cat.
USER: What kind of room does the couch appear to be in?
ASSISTANT: The couch appears to be in a living room setting.

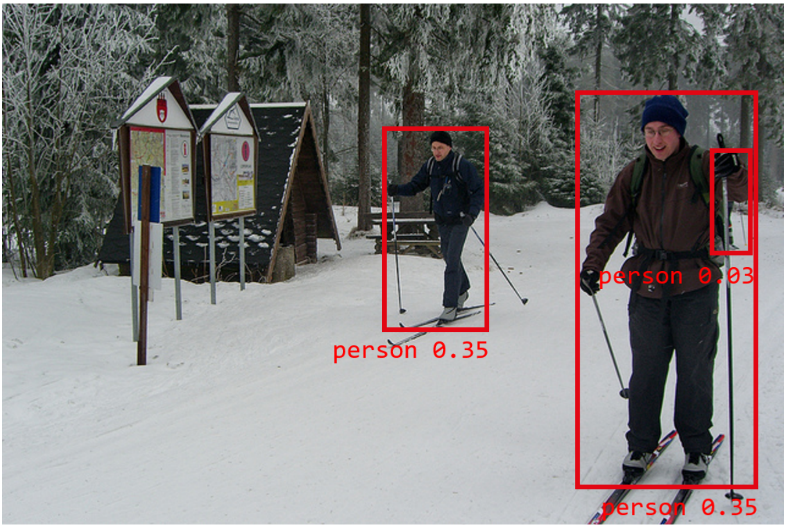
Original LLaVa:

USER: Where is the cat positioned in the image?
ASSISTANT: The cat is positioned on top of a couch in the image.
USER: What is the cat doing in the image?
ASSISTANT: The cat is sitting or standing on top of the couch, possibly looking out the window.
USER: What is the color of the curtains near the cat?
ASSISTANT: The curtains near the cat are red.
USER: Besides the cat, what else can be seen on the couch?
ASSISTANT: There are pillows on the couch, in addition to the cat.
USER: What kind of room does the couch appear to be in?
ASSISTANT: The couch appears to be in a living room.

$\mathcal{J}$:

USER: Where is the cat positioned in the image?
ASSISTANT: The cat is positioned on a couch in the image.
USER: What is the cat doing in the image?
ASSISTANT: The cat is sitting or laying on the couch.
USER: What is the color of the curtains near the cat?
ASSISTANT: The color of the curtains near the cat is red.
USER: Besides the cat, what else can be seen on the couch?
ASSISTANT: A blanket can also be seen on the couch along with the cat.
USER: What kind of room does the couch appear to be in?
ASSISTANT: The couch appears to be in a living room.

Figure 2: Model $\mathcal{J}$ answers in almost an identical fashion to both GT and LLaVa. However, instead of mentioning the pillows on the couch, the model mentions a blanket, which is also present and an equally valid response.



GT:

USER: Are the people in the image skiing downhill or cross-country skiing?
ASSISTANT: The people in the image are cross-country skiing in the woods, as they are skiing on a trail rather than a steep slope.
USER: How many people are in the image?
ASSISTANT: There are two people in the image, both on skis in the snow.
USER: What kind of environment are they skiing in?
ASSISTANT: They are skiing in a wooded environment, following a trail through the trees while surrounded by snow.
USER: Do the skiers have any additional gear with them besides their skis and poles?
ASSISTANT: Yes, the two male skiers are carrying backpacks while they ski through the woods. The backpacks might contain essentials for their skiing adventure, such as food, water, extra clothing, or safety equipment.

Original LLaVa:

USER: Are the people in the image skiing downhill or cross-country skiing?
ASSISTANT: The people in the image are cross-country skiing, as they are using skis and poles to move across the snow-covered ground.
USER: How many people are in the image?
ASSISTANT: There are two people in the image.
USER: What kind of environment are they skiing in?
ASSISTANT: The two people are skiing in a snowy environment, likely in a forest or a similar natural setting.
USER: Do the skiers have any additional gear with them besides their skis and poles?
ASSISTANT: Yes, the skiers have backpacks with them, which may contain essential items for their cross-country skiing adventure.

$\mathcal{J}$:

USER: Are the people in the image skiing downhill or cross-country skiing?
ASSISTANT: The people in the image are cross-country skiing, as they are standing on a snow-covered slope and holding their skis.
USER: How many people are in the image?
ASSISTANT: There are three people in the image.
USER: What kind of environment are they skiing in?
ASSISTANT: They are skiing in a snow-covered environment, which could be a mountain or a slope.
USER: Do the skiers have any additional gear with them besides their skis and poles?
ASSISTANT: Yes, the skiers have backpacks with them, which could be used to carry additional supplies or equipment for their cross-country skiing adventure.

Figure 3: Model $\mathcal{J}$ identifies three people in the image instead of two, as specified by the GT and in the LLaVa answer. However, as also indicated by $\mathbf{O}^{OWL}$, there is in fact a third person hidden behind the one on the right in foreground, as demonstrated by the barely visible ski and ski stick. We hypothesize that, since OWL-ViT is a more powerful detection model than the original LLaVa vision encoder, this information may have been lost with the original encoder. Instead, $\mathcal{J}$ was able to reason and answer also regarding this third person.

GT:

USER: What color is the dog in the image?
ASSISTANT: The dog in the image is brown.
USER: What type of collar does the dog have?
ASSISTANT: The dog has a red collar.
USER: What is the dog's position behind the window?
ASSISTANT: The dog is sitting behind the window.
USER: How does the dog appear to feel as it looks out the window?
ASSISTANT: As the dog stares out the window, it appears to feel longingly, perhaps suggesting a desire to be outside or interacting with someone or something that has caught its attention.
USER: Can you describe the dog's fur?
ASSISTANT: The dog has shaggy, fuzzy, and furry brown fur, which makes it look cute and cuddly.

Original LLaVa:

USER: What color is the dog in the image?
ASSISTANT: The dog in the image is brown.
USER: What type of collar does the dog have?
ASSISTANT: The dog has a leather collar.
USER: What is the dog's position behind the window?
ASSISTANT: The dog is sitting behind the window.
USER: How does the dog appear to feel as it looks out the window?
ASSISTANT: The dog appears to be feeling sad as it looks out the window.
USER: Can you describe the dog's fur?
ASSISTANT: The dog has long, shaggy fur.

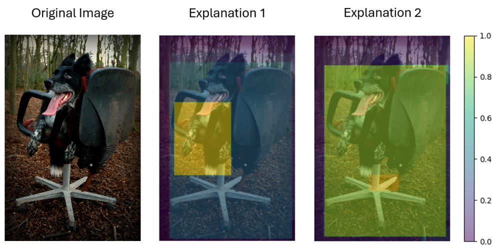
$\mathcal{J}$:

USER: What color is the dog in the image?
ASSISTANT: The dog in the image is white.
USER: What type of collar does the dog have?
ASSISTANT: The dog has a purple collar.
USER: What is the dog's position behind the window?
ASSISTANT: The dog is sitting behind the window.
USER: How does the dog appear to feel as it looks out the window?
ASSISTANT: The dog appears to be feeling sad as it looks out the window.
USER: Can you describe the dog's fur?
ASSISTANT: The dog has fluffy, white fur.

Figure 4: Model $\mathcal{J}$'s output differs from the GT and LLaVa in terms of colors of the dog and collar and in the qualities of the dog's fur. Nonetheless, these are reasonable responses. Indeed, the collar is purplish-red, the dog is a very light brown color, and its fur is indeed fluffy.

3 User Study Questions

We now present all the 10 questions that were asked to participants of the user study and report the number of correct responses for each question. These images also serve as additional examples of GA saliency maps.

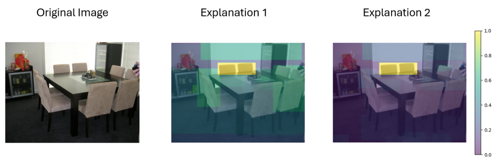


Q: Which explanation is for object "chair" and which is for object "dog"?

A1: Explanation 1 explains "chair" and Explanation 2 explains "dog"

A2: Explanation 1 explains "dog" and Explanation 2 explains "chair"

Correct answer is A2 with 100% correct answers.

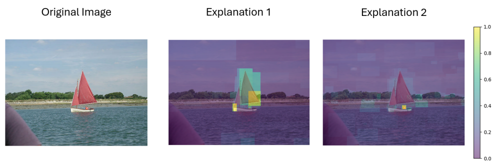


Q: Which explanation is for object "chair" and which is for object "table"?

A1: Explanation 1 explains "chair" and Explanation 2 explains "table"

A2: Explanation 1 explains "table" and Explanation 2 explains "chair"

Correct answer is A2 with 71% correct answers.



Q: Which explanation is for object "boat" and which is for object "person"?

A1: Explanation 1 explains "boat" and Explanation 2 explains "person"

A2: Explanation 1 explains "person" and Explanation 2 explains "boat"

Correct answer is A1 with 100% correct answers.

Q: Which explanation is for object "bicycle" and which is for object "person"?

A1: Explanation 1 explains "bicycle" and Explanation 2 explains "person"

A2: Explanation 1 explains "person" and Explanation 2 explains "bicycle"

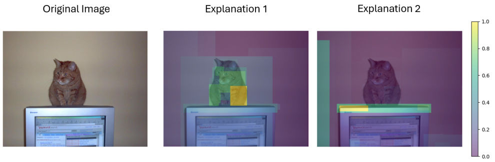
Correct answer is A1 with 100% correct answers.

Q: Which explanation is for object "person" and which is for object "chair"?

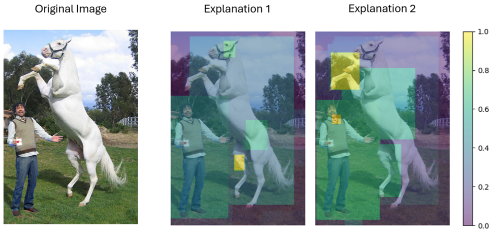
A1: Explanation 1 explains "person" and Explanation 2 explains "chair"

A2: Explanation 1 explains "chair" and Explanation 2 explains "person"

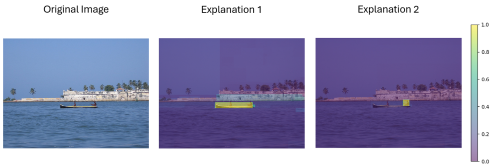
Correct answer is A2 with 100% correct answers.



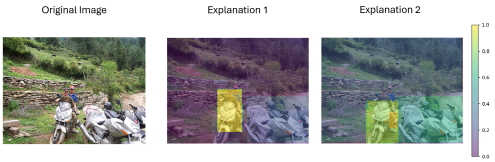
Q: Which explanation is for object "cat" and which is for object "TV/monitor"?
A1: Explanation 1 explains "cat" and Explanation 2 explains "TV/monitor"
A2: Explanation 1 explains "TV/monitor" and Explanation 2 explains "cat"
Correct answer is A1 with 100% correct answers.



Q: Which explanation is for object "person" and which is for object "horse"?
A1: Explanation 1 explains "person" and Explanation 2 explains "horse"
A2: Explanation 1 explains "horse" and Explanation 2 explains "person"
Correct answer is A2 with 71% correct answers.



Q: Which explanation is for object "person" and which is for object "boat"?
A1: Explanation 1 explains "person" and Explanation 2 explains "boat"
A2: Explanation 1 explains "boat" and Explanation 2 explains "person"
Correct answer is A2 with 100% correct answers.



Q: Which explanation is for object "person" and which is for object "motorbike"?
A1: Explanation 1 explains "person" and Explanation 2 explains "motorbike"
A2: Explanation 1 explains "motorbike" and Explanation 2 explains "person"
Correct answer is A1 with 100% correct answers.

Q: Which explanation is for object "person" and which is for object "horse"?
A1: Explanation 1 explains "person" and Explanation 2 explains "horse"
A2: Explanation 1 explains "horse" and Explanation 2 explains "person"
Correct answer is A1 with 100% correct answers.

4 Bias Assessment

As detailed in Section 4.4, we employ Stable Diffusion XL (SDXL) [4] to generate images and construct the portrait datasets used to assess model biases. We discuss generation details and give examples of the generated images. For the *biological gender* bias, we generate 50 portraits of female and 50 portraits of male people using prompt:

“A headshot of a <man|woman> in his 40s, DSLR, detailed, 8k, in perfect focus, shoulder height”

For the *ethnicity* bias, we generate 50 portraits of african-american and 50 portraits of caucasian people using prompt:

“A headshot of a <caucasian|african-american> person, DSLR, detailed, 8k, in perfect focus, shoulder height”

We use Stability AI’s official SDXL checkpoint `stable-diffusion-xl-base-1.0`, and run it using a diffuse-refine pipeline with 40 initial diffusion steps and 10 refinement steps. Generated images are available at: https://github.com/loris2222/ExplainingMLLMs.

Figure 5: Example images generated using each of the four prompts.

References

[1] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[3] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022.

[4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.