What Matters when Building Vision Language Models for Product Image Analysis?
— Supplementary Document —

Ameni Trabelsi[*] Maria Zontak[*] Yiming Qian[*] Brian Jackson Suleiman Khan Umit Batur
Amazon Inc.
Seattle, USA

This supplementary material provides additional results and examples comparing fine-tuned LLaVA (F-LLaVA) to proprietary models. Figure 1 qualitatively demonstrates F-LLaVA abilities to spatial ground product features in the image. Figures 2-4 showcase failure cases on the IBIE benchmark for Gemini, F-LLaVA, and a subjective example respectively. Figure 5 offers more details on Catalog-bench, a comprehensive benchmark of 440K tasks spanning 53 tasks, 44 product attributes, and 66 product types. Due to the non-uniform distribution of product types (top of Fig. 5), in the main paper we report performance as the average of averages per task. Figure 5 (bottom) demonstrates that F-LLaVA outperforms ChatGPT Turbo on all tasks.

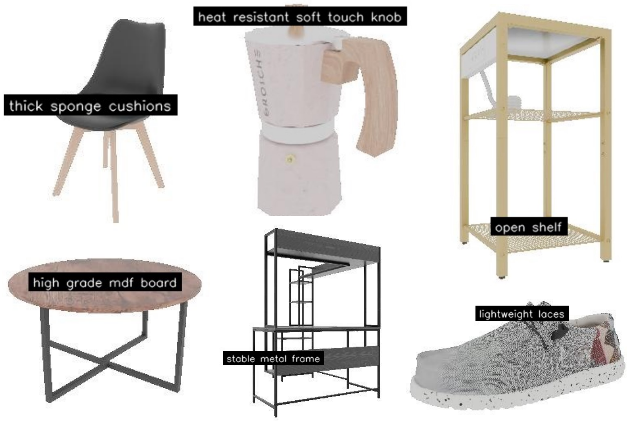


Figure 1 **Spatial Grounding Inference Examples**

---

[*] Authors contributed equally

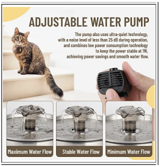
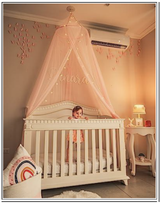
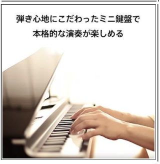
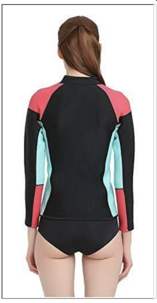
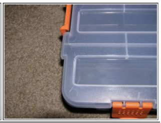
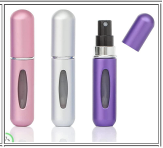
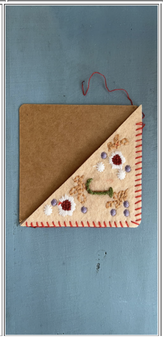
Feature	Expected Answer	Full Answer: FT_LLaVA	Short Answer: FT_LLaVA	Full Answer: Gemini	Short Answer: Gemini	Image
Water fountain design	Yes	{"reason": "The image shows an advertisement for an "Adjustable Water Pump" product. It displays a cat sitting on stairs, suggesting the pump is designed for use with pets like cats. The image also shows three different water flow patterns - maximum water flow, stable water flow, and minimum water flow.\n\nTo determine if the water fountain design can be inferred from the image, I need to look for any visual cues or information related to a water fountain design. The image does not explicitly show a water fountain, but it does depict the different water flow patterns that could potentially be used in a water fountain design.", "value": "Yes" }	Yes	{“reason”: “While the image showcases the water fountain pump and its adjustable flow settings, it doesn't reveal the fountain's overall design or structure. We can't infer if it's a tiered fountain, one with a bowl, or any other particular design." , “value”: “B”} \n	No	 ADJUSTABLE WATER PUMP The pump also uses ultra-quiet technology, with a noise level of less than 25 dB during operation, and combines low power consumption technology to keep the power stable at 1W, achieving power savings and smooth water flow. Maximum Water Flow Stable Water Flow Minimum Water Flow
2 star lights hanging from canopy	Yes	{"reason": "The image shows a nursery or baby's room with a white crib, a pink canopy with strings of lights hanging from it, and a pink curtain with the name "Dinara" written on it. There is also a white dresser or cabinet with a lamp on top.\n\nTo determine if the 2 star lights hanging from the canopy can be inferred from the image, I will carefully examine the canopy and look for any star-shaped lights or decorations.", "value": "Yes" }	Yes	{“reason”: “The image shows a canopy hanging over a crib, draped over a circular frame. While the frame obscures the exact attachment point of the lights, it's unlikely that they hang independently from the canopy itself." , “value”: “B”} \n	No	
Black and white keys	Yes	{"reason": "The image shows a close-up view of a person's hands playing a piano keyboard. The keys on the keyboard are clearly visible, with some keys appearing black and others appearing white. The text at the top of the image mentions "音楽" which means "music" in Japanese.\n\nTo determine if the black and white keys can be inferred from the image, I need to carefully examine the keyboard and see if there is any information physically present that confirms the presence of both black and white keys.", "value": "Yes" }	Yes	{"reason": "Although the image shows a portion of the keyboard, it is not clear enough to determine if the keys are black and white.", "value": "B"}	No	
High neck collar	Yes	{"reason": "The image shows a woman wearing a black wetsuit or rash guard with colorful panels on the shoulders and chest area. The wetsuit has a high neck collar that covers the neck area. Based on this visual information, I can infer that the wetsuit has a high neck collar design.", "value": "Yes" }	Yes	{“reason”: “The collar of the product is not visible in the image, her hair is covering it. ”, “value”: “B”}	No	
Product type: Tool box organizer	No	{"reason": "The image shows a clear plastic container with multiple compartments. It appears to be a storage container or organizer for small items. The container has an orange handle on the top, suggesting it is portable and can be carried around.\n\nTo determine if this product can be inferred as a "tool box organizer" from the information in the image, I will examine the visual details:\n\n- The container has multiple compartments, which could be useful for organizing tools or small parts.\n- The orange handle indicates it is designed to be carried, which aligns with a tool box organizer.\n- However, the image does not explicitly show any tools or mention the product being specifically for tools.\n- The container appears to be made of clear plastic, which is a common material for tool box organizers, but not definitive.", "value": "No" }	No	{“reason”: “The image shows a plastic container with multiple small compartments and a latch. This is indicative of a tool box organizer.”, “value”: “A”}	Yes	
Aluminum spray bottles	No	{"reason": "The image shows three cylindrical objects that appear to be spray bottles or containers. They are made of a metallic material, likely aluminum or a similar metal alloy. The objects have a pink/rose gold color on the top portion and a purple/mauve color on the bottom portion.\n\nTo determine if these are aluminum spray bottles, I need to look for any explicit information or labels on the objects that indicate their material composition. However, I don't see any text or markings on the objects themselves that confirm they are made of aluminum.\n\nWhile the metallic appearance and cylindrical shape are consistent with spray bottles, there is no definitive evidence in the image that confirms they are specifically aluminum spray bottles. It's possible they could be made of another metal or material.", "value": "No" }	No	{“reason”: “The image shows three spray bottles of the same design but with different colors, one of them is silver. Aluminum is a light metal with a silver color, so it is possible that the silver bottle is made of aluminum.”, “value”: “A”}	Yes	
Can be used as page markers, book decorations, jewelry making, DIY crafts, or wedding gifts	No	{"reason": "The image shows a decorative item made of cardboard or paper. It has a triangular shape with a red string or cord attached to it. The item is decorated with various materials like beads, feathers, and other embellishments in colors like red, purple, and green.\n\nGiven the decorative and crafty nature of this item, it could potentially be used for various purposes like page markers, book decorations, jewelry making, DIY crafts, or even as a wedding gift. However, the image itself does not explicitly state or confirm any of these intended uses.", "value": "No" }	No	{“reason”: “The image shows a decorated piece of card. It is triangular in shape and decorated with floral embroidery. This type of product could be used as a bookmark, for book decoration, in DIY crafts, or as a wedding gift. There is no indication it could be used for jewelry making.”, “value”: “A”}	Yes	

FIGURE 2 **IBIE Benchmark, F-LLaVA vs. Gemini Pro**: The reasoning and the answer is presented for F-LLAVA and Gemini-Pro. The reason can help us to understand the decision making process of a model for debugging purposes. In these examples, the answer of *F-LLaVA agrees with human label (expected answer).*

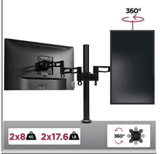
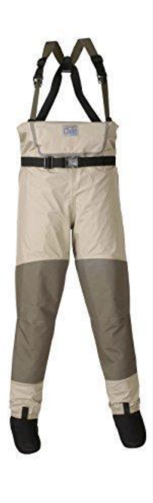
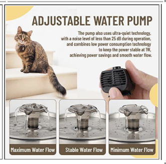
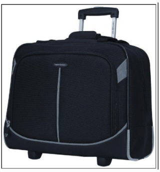
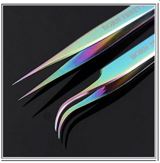
Feature	Expected Answer	Full Answer: FT_LLaVA	Short Answer: FT_LLaVA	Full Answer: Gemini	Short Answer: Gemini	Image
Hold monitors up to 8kg each	Yes	{"reason": "The image shows a monitor mount or stand with a 360-degree rotating arm. It appears to be designed to hold a monitor or display. The mount has a weight capacity of 28kg (61.6 lbs) according to the information provided in the image.\n\nTo determine if the mount can hold monitors up to 8kg each, I need to consider the weight capacity and the typical weight of monitors. Most modern monitors weigh between 1-3 kg (2-6 lbs). The weight capacity of 28kg (61.6 lbs) seems sufficient to hold two monitors of typical weight, but it's not explicitly stated that it can accommodate two monitors.", "value": "No" }	No	{“reason”: “The text states ‘2x8 KG’ clearly indicating each monitor arm supports up to 8 KG”, “value”: “A”}	Yes	
Reinforced leg panels	No	{"reason": "The image shows a pair of tan-colored pants or trousers made of a lightweight, breathable material. The pants have reinforced knee panels made of a darker, more durable fabric. There are also reinforced panels on the lower legs, likely to provide protection from abrasions or impacts. The pants have suspender-style straps attached to the sides, suggesting they are designed for activities like hiking or outdoor work where extra support and coverage is needed. Based on the visible reinforced panels on the knees and lower legs, it appears these pants are intended to provide protection and durability for the wearer's legs.", "value": "Yes" }	Yes	{"reason": "The image does not provide a clear enough view of the leg panels to determine if they are reinforced.", "value": "B"}	No	
Water pump runs below 25dB	Yes	{"reason": "The image shows an advertisement for an "Adjustable Water Pump" product. It displays a cat sitting on stairs, suggesting the pump is designed for use with pets. The image highlights three key features of the pump:\n\n1. "Maximum water flow" - This likely refers to the pump's ability to produce a high volume of water flow.\n\n2. "Stable water flow" - This suggests the pump maintains a consistent water flow rate.\n\n3. "Minimum water flow" - This implies the pump can operate at a low water flow rate.\n\nHowever, I do not see any specific information about the noise level or decibel rating of the pump. The image focuses more on the water flow capabilities rather than acoustic performance.", "value": "No" }	No	```json\n{\n "reason": "The description states: \"The pump also uses ultra-quiet technology, with a noise level of less than 25 dB during operation\".",\n "value": "A"\n}\n```	Yes	 ADJUSTABLE WATER PUMP Maximum Water Flow Stable Water Flow Minimum Water Flow
Inline skate wheels	No	{"reason": "The image shows a black rolling suitcase or travel bag with wheels. The wheels appear to be small, round wheels that allow the bag to roll on its base. Based on the visible wheels in the image, I can infer that this bag has inline skate wheels for mobility.", "value": "Yes" }	Yes	{"reason": "The image shows a rolling suitcase, and there is no information about its wheels being related to inline skates. ", "value": "B"}	No	
Stainless steel material	No	{"reason": "The image shows three metallic objects that appear to be tweezers or precision tools. They have a curved, pointed shape with a flat base. The objects are made of a shiny, reflective material that has a rainbow-like iridescent coloring.\n\nTo determine if the material is stainless steel, I need to look for any explicit mention or labeling of the material used. The image does not provide any text or labels indicating the specific material composition.\n\nHowever, based on the appearance and typical use of such precision tools, stainless steel would be a reasonable inference. Stainless steel is a common material used for tools and instruments due to its durability, corrosion resistance, and ability to maintain a polished, reflective surface.", "value": "Yes" }	Yes	{“reason”: “Although the tweezers have a rainbow finish, this does not definitively indicate the underlying material. The finish could be applied to various metals." , “value”: “B”}	No	

Figure 3 **IBIE Benchmark, F-LLaVA vs. Gemini Pro**: The reasoning and the answer is presented for F-LLAVA and Gemini-Pro. The reason can help us to understand the decision making process of a model for debugging purposes. In these examples, the answer of *Gemini agrees with human label (expected answer).*

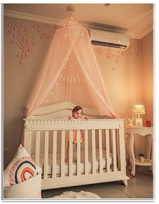
Feature	Expected Answer	Full Answer: FT_LLaVA	Short Answer: FT_LLaVA	Full Answer: Gemini	Short Answer: Gemini	Image
Bed canopy for girls	Yes	{"reason": "The image shows a nursery or baby's room with a white crib or baby bed. Above the crib is a pink canopy or tent-like structure with strings of lights hanging from it. The canopy appears to be designed for a baby girl's room.\n\nTo determine if the bed canopy is specifically for girls, I need to look for any explicit indications or text in the image that confirms this. The pink color and decorative style of the canopy suggest it is intended for a girl's room, but this is an inference and not definitively stated.", "value": "No" }	No	{"reason": "The pink color scheme, butterflies on the wall, and the name \"Dinara\" (a predominantly female name) on the canopy suggest the bed canopy is intended for a girl.", "value": "A"}	Yes	

Figure 4 **IBIE Benchmark, F-LLaVA vs. Gemini Pro**: This example illustrates that sometimes the expected answer is subjective and in fact both models can be right following certain reasoning

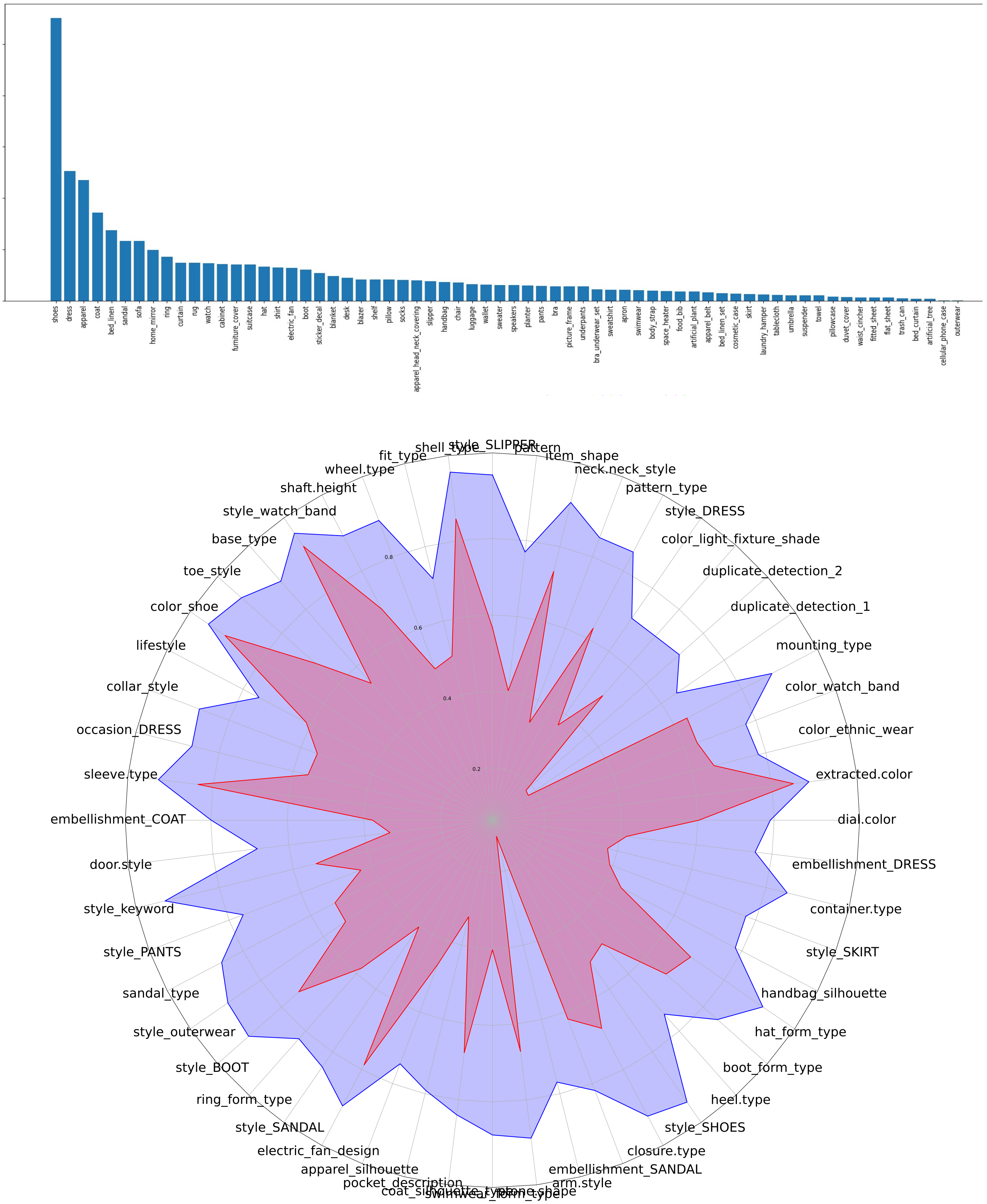


FIGURE 5 **Additional details of Catalog-bench:** Top: product type distribution Bottom: F-LLaVA (blue) vs. GPT4 Turbo (red)