Part I

Appendix

Table of Contents

A User Interface **18**

B Question Category and Image Domain **19**

C Analysis of Failure Cases **22**

D Data Analysis **28**

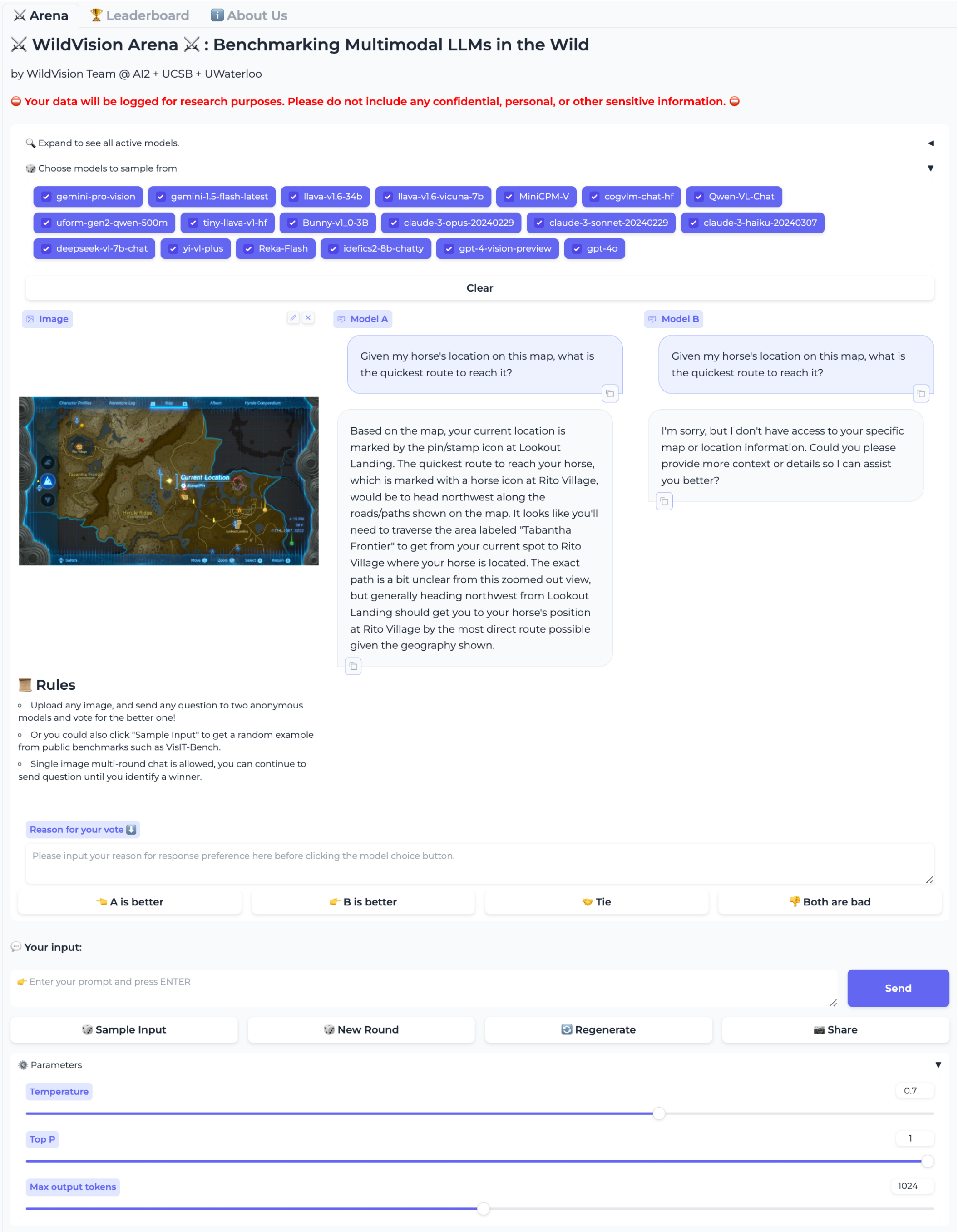
E Prompt Template **29**
- E.1 Taxonomy Annotation 29
- E.2 VLM Voting 30
- E.3 WILDVISION-BENCH Evaluator 31

F Discussions **32**
- F.1 Limitations 32
- F.2 Societal Impact 32

G Accessiblity of Datasets **32**
- G.1 Dataset Documentation and Intended Uses 32
- G.2 Maintenance Plan 32
- G.3 Author Statement 32

A User Interface

In Figure 8, we show a screenshot of the user interface of our WildVision-Arena, which presents an interactive environment for evaluating multimodal large language models. This environment allows users to input questions and compare responses from multiple models simultaneously. Each model's answer is displayed side-by-side, enabling a straightforward comparison of their performance and capabilities based on user queries related to specific images or tasks. The interface also facilitates easy selection and voting to decide which model's response fits the user's criteria best, enhancing the user's ability to judge and refine the models' outputs effectively.


⚔ Arena
🏆 Leaderboard
ℹ About Us

⚔ WildVision Arena ⚔ : Benchmarking Multimodal LLMs in the Wild

by WildVision Team @ AI2 + UCSB + UWaterloo

⛔ Your data will be logged for research purposes. Please do not include any confidential, personal, or other sensitive information. ⛔

🔍 Expand to see all active models.
🎲 Choose models to sample from

☑ gemini-pro-vision ☑ gemini-1.5-flash-latest ☑ llava-v1.6-34b ☑ llava-v1.6-vicuna-7b ☑ MiniCPM-V ☑ cogvlm-chat-hf ☑ Qwen-VL-Chat
☑ uform-gen2-qwen-500m ☑ tiny-llava-v1-hf ☑ Bunny-v1_0-3B ☑ claude-3-opus-20240229 ☑ claude-3-sonnet-20240229 ☑ claude-3-haiku-20240307
☑ deepseek-vl-7b-chat ☑ yi-vl-plus ☑ Reka-Flash ☑ idefics2-8b-chatty ☑ gpt-4-vision-preview ☑ gpt-4o

Model A

Given my horse's location on this map, what is the quickest route to reach it?

Based on the map, your current location is marked by the pin/stamp icon at Lookout Landing. The quickest route to reach your horse, which is marked with a horse icon at Rito Village, would be to head northwest along the roads/paths shown on the map. It looks like you'll need to traverse the area labeled "Tabantha Frontier" to get from your current spot to Rito Village where your horse is located. The exact path is a bit unclear from this zoomed out view, but generally heading northwest from Lookout Landing should get you to your horse's position at Rito Village by the most direct route possible given the geography shown.

Model B

Given my horse's location on this map, what is the quickest route to reach it?

I'm sorry, but I don't have access to your specific map or location information. Could you please provide more context or details so I can assist you better?

📜 Rules

- Upload any image, and send any question to two anonymous models and vote for the better one!
- Or you could also click "Sample Input" to get a random example from public benchmarks such as VisIT-Bench.
- Single image multi-round chat is allowed, you can continue to send question until you identify a winner.

Temperature 0.7

Top P 1

Max output tokens 1024

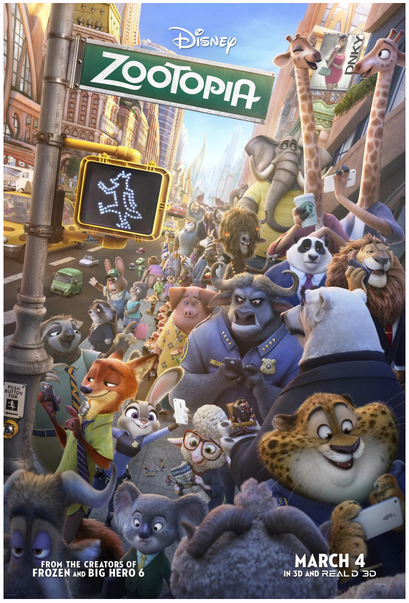
Figure 8: User Interface of WildVision-Arena.

B Question Category and Image Domain

In Table 6- 8, we showcase example data under each of the image domain and question category from WILDVISION-ARENA's users.

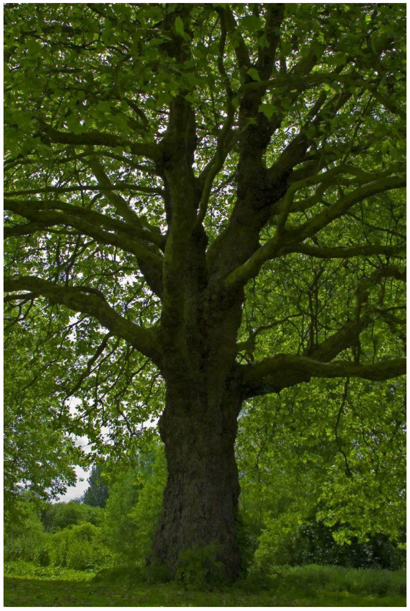
Table 6: Example input data in WILDVISION-ARENA tagged with `[Image Domain-Subdomain]` and `[ Question Category-Subcategory]`.

Image `[Entertainment-Movies/TV Shows]`



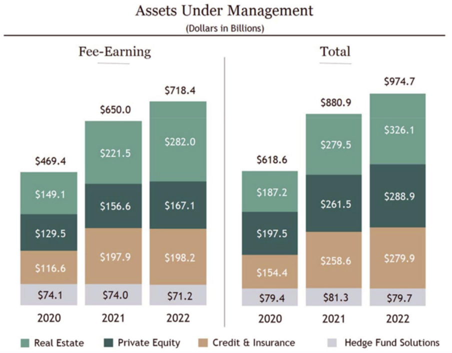
`[Descriptive-Movies/TV Shows]` **Text Prompt:** What are the two giraffe characters on this movie poster doing?

Image `[Natural-Plants]`



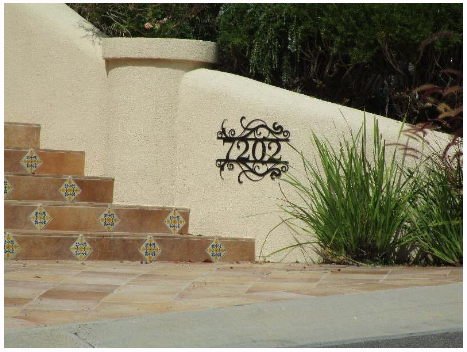
`[Analytical-Problem Solving]` **Text Prompt:** How likely is it to snow after this picture was taken? What would change with this type of tree before it's likely to snow?

Image `[Expert-Business]`



`[Analytical-Data Analysis]` **Text Prompt:** Which of the companies featured in the dashboard are headquartered outside the US?

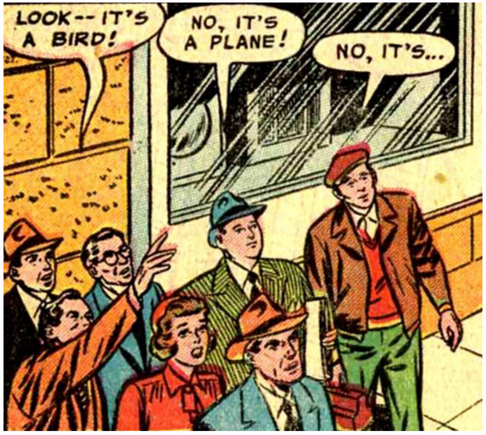
Image `[Urban-Infrastructure]`



`[Recognition-Text]` **Text Prompt:** Can you tell me the potential risks and the unreasonale parts in the image?

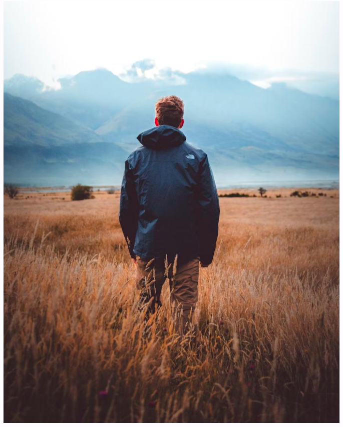
Table 7: Example input data in WILDVISION-ARENA tagged with `[Image Domain-Subdomain]` and `[ Question Category-Subcategory]`.

Image `[Entertainment-Comics]`



`[Descriptive-Scene Description]` **Text Prompt:** Whos's in the sky?

Image `[People-Portraits]`



`[Creative-Media Post]` **Text Prompt:** write a social media post with the provided image, saying that I am ready for the new challange.

Image `[Urban-Buildings]`

`[Recognition-Location]` **Text Prompt:** where is this?

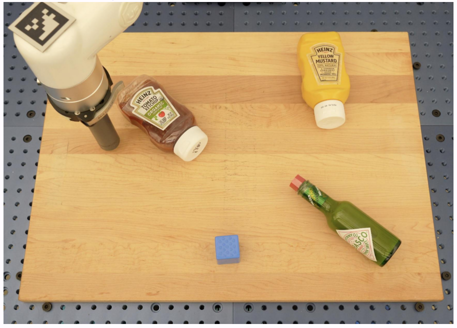
Image `[Expert-Science]`

`[Analytical-Safety Procedures]` **Text Prompt:** Can you tell me the potential risks and the unreasonale parts in the image?

Image `[Natural-Landscapes]`

`[Recognition-Location]` **Text Prompt:** where was this photo taken?

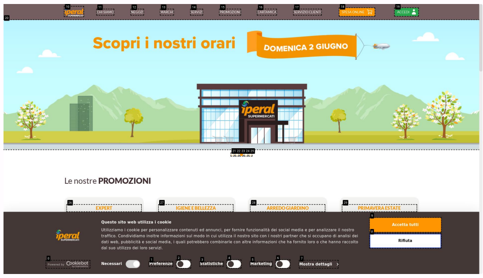
Image `[Objects-Household Tools]`



`[Descriptive-Object Description]` **Text Prompt:** describe the scene and objects

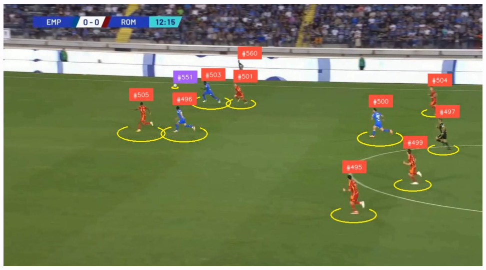
Table 8: Example input data in WILDVISION-ARENA tagged with `[Image Domain-Subdomain]` and `[ Question Category-Subcategory]`.

Image `[Entertainment-Web and Mobile Apps Screenshots]`



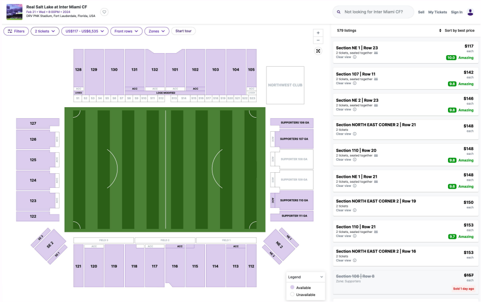
`[Interactive-Web Navigation]` **Text Prompt:** I need to download flyer, you will be given screenshot from browser with elements marked with number. give next action to take on web page to download the flyersngive me response in below format example 1 action:[click,scroll,wait], box:1 format action:, box:

Image `[Event-Sports]`



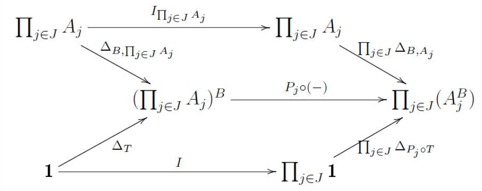
`[Descriptive-Scene Description]` **Text Prompt:** this is a football match , every player has an identifier , describe every player action (example : player #501 is running)

Image `[Urban-Infrastructure]`



`[Interactive-Recommendations]` **Text Prompt:** Which section's ticket would you recommend I purchase?

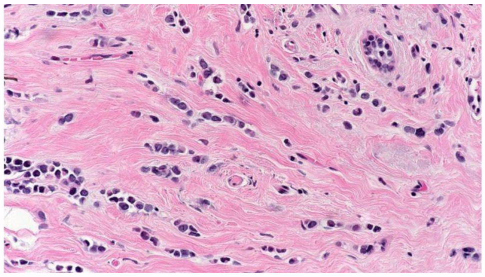
Image `[Expert-Science]`



$$\begin{tikzcd}
\prod_{j\in J} A_j \arrow[rr, "I_{\prod_{j\in J} A_j}"] \arrow[dr, "\Delta_{B,\prod_{j\in J} A_j}"] & & \prod_{j\in J} A_j \arrow[dr, "\prod_{j\in J} \Delta_{B,A_j}"] & \\
& (\prod_{j\in J} A_j)^B \arrow[rr, "P_j\circ(-)"] & & \prod_{j\in J}(A_j^B) \\
\mathbf{1} \arrow[ur, "\Delta_T"'] \arrow[rr, "I"] & & \prod_{j\in J} \mathbf{1} \arrow[ur, "\prod_{j\in J} \Delta_{P_j\circ T}"'] &
\end{tikzcd}$$

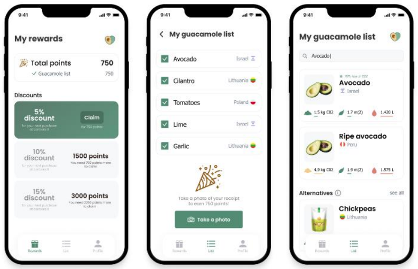
`[Interactive-Code Generation]` **Text Prompt:** Give me Latex code to create this diagram

Image `[Expert-Health and Medcine]`



`[Recognition-Object]` **Text Prompt:** what type of tumor is this?

Image `[Entertainment-Web and Mobile Apps Screenshots]`



`[Analytical-Critical Reviews]` **Text Prompt:** Review each screenshot carefully, focusing on different aspects of usability...

C Analysis of Failure Cases

We observe some common failure patterns of VLMs in the wild from WILDVISION-ARENA chat data. In Tables 9- 13, we present specific failure cases. Based on the types of errors, we have condensed six categories, detailed in the following paragraphs.

Visual Recognition Failures in this category involve several types of recognition challenges. Complex text, such as artistic Chinese characters (Error *#5* in Table 10), and small details, such as the text on a shop sign in a restaurant scene (Error *#3* in Table 9), often elude accurate detection. Errors also occur with small text in screenshots (Error *#18* in Table 13), recognizing objects against novel backgrounds (Error *#9* in Table 11), and identifying new objects within known contexts, like a helmet next to a motorcycle (Error *#11* in Table 11). Failures in interpreting charts (Error *#4* in Table 9) and identifying reflections (Error *#20* in Table 13) are also noted.

Visual Reasoning This category covers the ability to discern visual differences and reason through visual data. Errors include overlooking changes in clothes of the character (Error *#6* in Table 10), strategizing in chess (Error *#1* in Table 9), and reading analog clocks (Error *#16* in Table 11).

Spatial Imagination Challenges in this area involve imagining the outcomes of spatial transformations. Failures are evident in cube folding tasks (Error *#10* in Table 11) and visual riddles involving the folding of objects like magazines (Error *#12* in Table 11).

Contextual Understanding Errors in this category relate to recognizing and understanding contexts specific to certain domains, such as movies (Error *#7* in Table 10) and video games (Error *#8* in Table 10). Failures also include responding inappropriately to marked images (Error *#17* in Table 13).

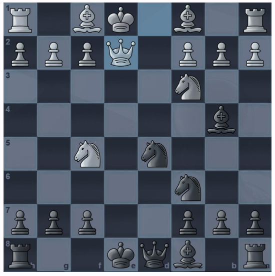
Expert Domain Knowledge This involves specialized knowledge areas where VLMs fail to provide accurate information or analysis, such as verifying architectural compliance (Error *#2* in Table 9) and analyzing programming code details (Error *#15* in Table 11).

Hallucination Errors of hallucination occur when models generate incorrect or misleading content based on subtle cues or nonexistent details, such as fictitious characters (Error *#13* in Table 12) or obscure brand names (Error *#18* in Table 13).

Safety Issues in this category are critical as they involve handling harmful, biased, or inappropriate content, and dealing with NSFW images in ways that either excessively censor or insufficiently filter content. These images are not presented and are filtered out in benchmark curation.

Table 9: Failure Cases.

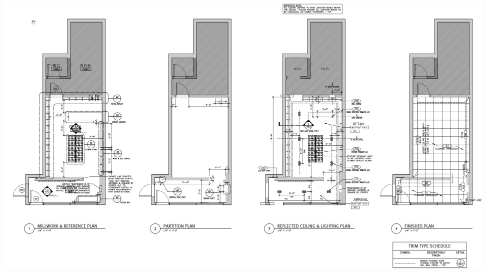
Image Error *#1*



Text Prompt: What should I play in this position?

GPT-4V: This chess position is unconventional and humorous because it features a board filled with knights (horses) except for a single queen. In a standard game of chess, this situation would not occur, as each player starts with just two knights. In this specific position, there's not much strategy to discuss because the game doesn't reflect an actual playable scenario. However, if we want to engage with this as a fun exercise in imagination, you could consider moving one of the knights that isn't pinned or creating a threat to the queen. For example, you could move one of the knights on the third rank to attack the queen, such as Nc3-d5 or Ne3-c4. Remember, this is just a playful scenario and not representative of a legitimate chess position. In a real game, the best move would depend on a much broader context of the board and the strategic positions of all the pieces, not just an army of knights.

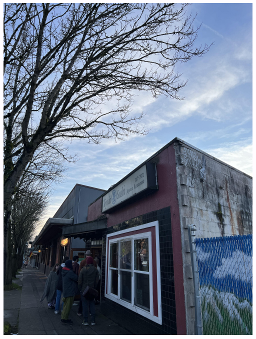
Image Error *#2*



Text Prompt: Based on the architecture plans is this compliant in the Ontario building code

GPT-4V: I'm sorry, but as an AI, I do not have the capacity to verify the compliance of architectural plans with specific building codes. Building code compliance typically involves a wide range of considerations, including but not limited to accessibility, fire safety, structural integrity, and mechanical systems. Professional architects and engineers, often alongside local code enforcement officials, ensure that designs and construction meet the appropriate codes and regulations. To determine compliance with the Ontario Building Code, the plans would need to be reviewed by a certified professional who has access to and understands the local regulations and requirements. They would check if the design meets all necessary stipulations for things such as room sizes, egress requirements, bathroom accessibility, fire ratings, and many other details. You should consult with a licensed Ontario architect or a building code consultant for an official assessment of compliance.

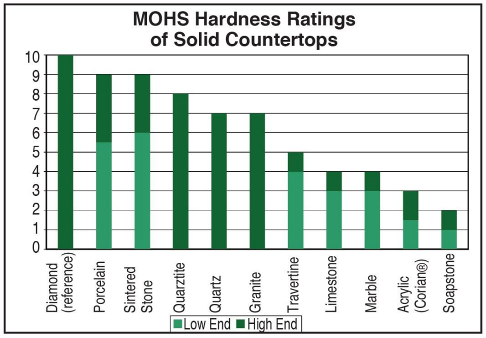
Image Error *#3*



Text Prompt: What food will these people have very soon?

LLaVA-NEXT-34B: Based on the image, the people are lined up outside a building with a sign that reads "Salsa's Authentic Mexican Restaurant." It seems likely that they will be having Mexican food very soon, which could include items such as tacos, burritos, enchiladas, chiles rellenos, guacamole, and other similar dishes that are typically served at Mexican restaurants. The specific menu items will depend on the restaurant's offerings and the preferences of the customers.

Image Error *#4*

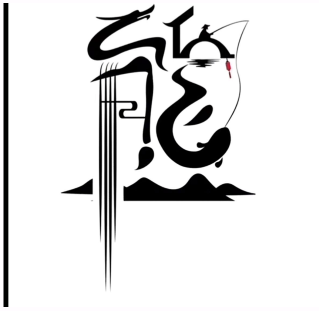


Text Prompt: Based on this chart, what is the MOHS hardness of Marble?

LLaVA-NEXT-34B: The MOHS hardness of Marble, as indicated on the chart, is 3.

Table 10: Failure Cases.

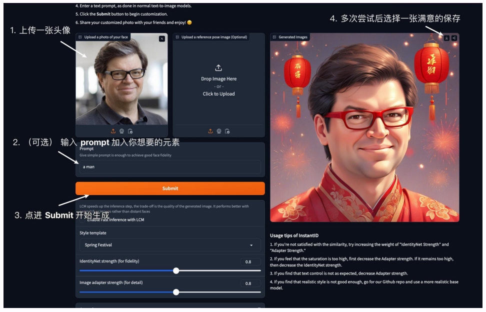
Image Error #5



Text Prompt: 图片中是什么字

GPT-4V: 图片中的字体设计巧妙地融合了艺术与文字，呈现出既像文字又像艺术作品的效果。这个特殊的设计使得文字的具体含义不太直观，没有提供清晰的线索来解释这些字符的标准解读。如果你有更多关于这些字的背景信息或上下文，我可以试图进行更多的解释。不过，从图像中直接解读标准文本是非常困难的。

Image Error #6



4. Enter a text prompt, as done in normal text-to-image models.
5. Click the **Submit** button to begin customization.
6. Share your customized photo with your friends and enjoy! 😊

1. 上传一张头像

2. （可选） 输入 **prompt** 加入你想要的元素

3. 点进 **Submit** 开始生成

4. 多次尝试后选择一张满意的保存

Upload a photo of your face
Upload a reference pose image (Optional)
Drop Image Here
- or -
Click to Upload
Generated Images
Submit

Prompt
Give simple prompt is enough to achieve good face fidelity
a man

LCM speeds up the inference step, the trade-off is the quality of the generated image. It performs better with close-up face images rather than distant faces
Enable Fast Inference with LCM

Style template: Spring Festival

IdentityNet strength (for fidelity): 0.8

Image adapter strength (for detail): 0.8

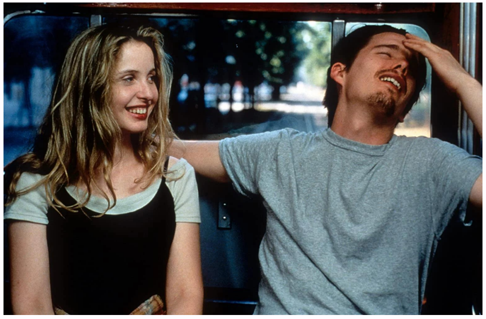
Usage tips of InstantID

1. If you're not satisfied with the similarity, try increasing the weight of "IdentityNet Strength" and "Adapter Strength."
2. If you feel that the saturation is too high, first decrease the Adapter strength. If it remains too high, then decrease the IdentityNet strength.
3. If you find that text control is not as expected, decrease Adapter strength.
4. If you find that realistic style is not good enough, go for our Github repo and use a more realistic base model.

Text Prompt: 这张图里右侧的照片和左侧相比有什么变化

Gemini-Pro-Vision: 右侧的照片中，人物的脸部表情更加丰富，并且背景中添加了春节的元素，使画面更具喜庆的氛围。

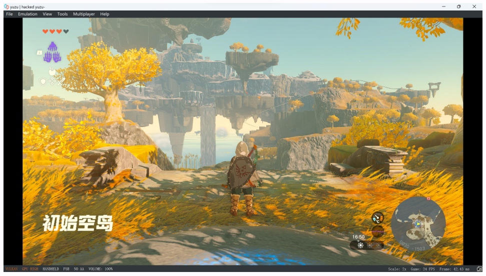
Image Error #7



Text Prompt: which movie is this picture from? what is it about?

LLaVA-NEXT-34B: This image appears to be from the movie T̈he Girl(̈2012), which is a psychological thriller film directed by David Riker and starring Abbie Cornish and Caleb Landry Jones. The film is about a young woman who is hired to care for a girl living in a remote Texas farmhouse, but soon discovers that the girl is not what she seems. The story revolves around themes of isolation, identity, and the consequences of one's actions.

Image Error #8

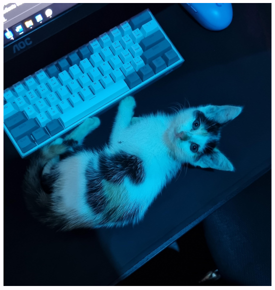


Text Prompt: do u know which game i am playing?

LLaVA-NEXT-34B: You are playing T̈he Legend of Zelda: Breath of the Wild,ẅhich is an action-adventure game developed and published by Nintendo for the Wii U and Nintendo Switch consoles. The game is known for its open-world exploration and its unique gameplay mechanics.

Table 11: Failure Cases.

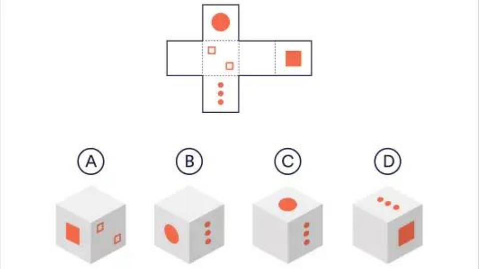
Image Error *#9*



Text Prompt: How many keys does the keyboard have?

GPT-4V: The keyboard in the image features a non-standard layout, which appears to be a compact form factor lacking the numeric keypad commonly found on full-sized variants. Without being able to view the entire keyboard, it's impossible to provide an exact count of the keys...

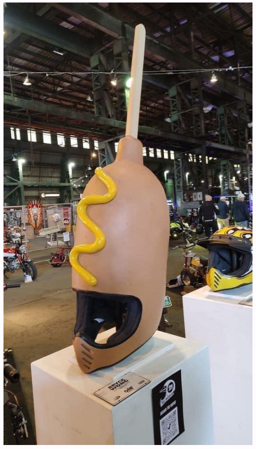
Image Error *#10*



Text Prompt: By examining the unfolded cube (net) at the top of the image, which one of the four cubes labeled A, B, C, and D can be created by folding this net along the edges?

Gemini-Pro-Vision: The correct answer is C. The unfolded cube (net) at the top of the image has two red squares and three orange circles. Cube C has two red squares and three orange circles in the same arrangement as the unfolded cube.

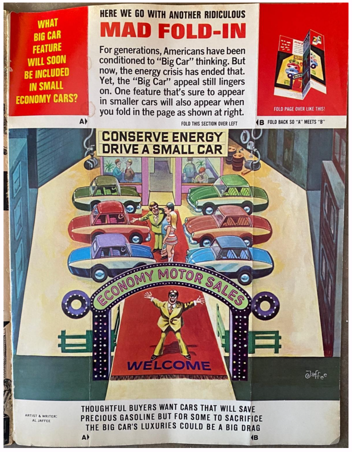
Image Error *#11*



Text Prompt: what is this

GPT-4V: This is an oversized, artistic representation of a classic hand-held manual egg beater or whisk. It's designed to mimic the appearance of a traditional egg beater with an exaggerated size, likely for decorative or exhibit purposes as suggested by the setting that appears to be a show or exhibition hall. The inclusion of this piece in a public space may indicate it being part of an art installation, event display, or perhaps a promotion related to culinary equipment or innovation.

Image Error *#12*



WHAT BIG CAR FEATURE WILL SOON BE INCLUDED IN SMALL ECONOMY CARS?

HERE WE GO WITH ANOTHER RIDICULOUS

MAD FOLD-IN

For generations, Americans have been conditioned to "Big Car" thinking. But now, the energy crisis has ended that. Yet, the "Big Car" appeal still lingers on. One feature that's sure to appear in smaller cars will also appear when you fold in the page as shown at right.

FOLD PAGE OVER LIKE THIS!

A▸ FOLD THIS SECTION OVER LEFT ◂B FOLD BACK SO "A" MEETS "B"

ARTIST & WRITER: AL JAFFEE

THOUGHTFUL BUYERS WANT CARS THAT WILL SAVE PRECIOUS GASOLINE BUT FOR SOME TO SACRIFICE THE BIG CAR'S LUXURIES COULD BE A BIG DRAG

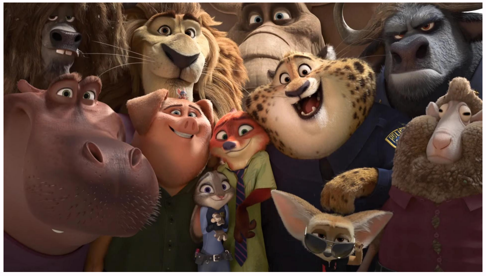
A▸ ◂B

Text Prompt: Try to guess the solution to this Fold-in from Mad magazine. A Mad Fold-in is a visual riddle that a human can easily solve by physically folding the page as explained in the instructions in the top-right corner. I want to verify if you are able to solve it. What big car feature will soon be included in small economy cars?

LLaVA-NEXT-34B: The Mad Fold-in features a car salesman in front of a sign that says "CONSERVE ENERGY DRIVE A SMALL CAR" with various "big car" features like a tire iron, a spare tire, a picnic basket, and a trunk full of items. When you fold the page as directed, it's revealed that the large "big car" features are actually inside the small car, showing that the small car also has these items, humorously implying that the small car will soon include all the features typically associated with big cars. The joke is that despite their size, small cars will also include all the amenities that were once exclusive to larger vehicles.

Table 12: Failure Cases.

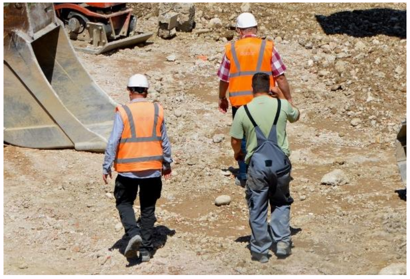
Image Error *#13*



Text Prompt: What is the name of the character with the sunglasses?

GPT-4V: You're looking at an image from the animated film Z̈ootopia.T̈he character with the sunglasses is named Nick Wilde; he's a red fox who is one of the main characters in the movie.

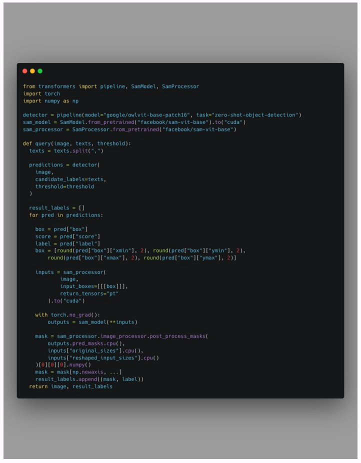
Image Error *#14*



Text Prompt: Are the workers wearing safety devices?

LLaVA-NEXT-34B: In the image you've provided, the workers are wearing safety vests, which are commonly used in construction and industrial environments to increase visibility and reduce the risk of accidents. These vests are a standard safety device to help workers be more visible to operators of heavy machinery and other workers in bright colors that contrast with the surroundings. The workers also appear to be wearing hard hats, which are required on construction sites to protect the head from potential injuries.

Image Error *#15*

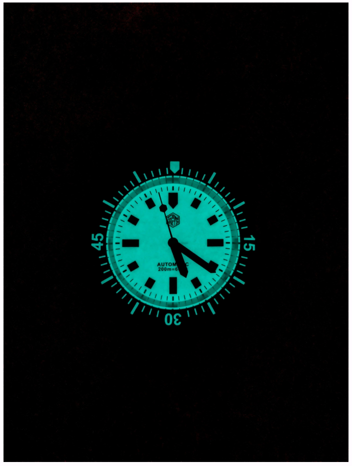


```
from transformers import pipeline, SamModel, SamProcessor
import torch
import numpy as np

detector = pipeline(model="google/owlvit-base-patch16", task="zero-shot-object-detection")
sam_model = SamModel.from_pretrained("facebook/sam-vit-base").to("cuda")
sam_processor = SamProcessor.from_pretrained("facebook/sam-vit-base")

def query(image, texts, threshold):
  texts = texts.split(",")

  predictions = detector(
    image,
    candidate_labels=texts,
    threshold=threshold
  )

  result_labels = []
  for pred in predictions:

    box = pred["box"]
    score = pred["score"]
    label = pred["label"]
    box = [round(pred["box"]["xmin"], 2), round(pred["box"]["ymin"], 2),
        round(pred["box"]["xmax"], 2), round(pred["box"]["ymax"], 2)]

    inputs = sam_processor(
            image,
            input_boxes=[[[box]]],
            return_tensors="pt"
        ).to("cuda")

    with torch.no_grad():
        outputs = sam_model(**inputs)

    mask = sam_processor.image_processor.post_process_masks(
        outputs.pred_masks.cpu(),
        inputs["original_sizes"].cpu(),
        inputs["reshaped_input_sizes"].cpu()
    )[0][0][0].numpy()
    mask = mask[np.newaxis, ...]
    result_labels.append((mask, label))
  return image, result_labels
```

Text Prompt: What do you see in the image?

Gemini-Pro-Vision: This is a code snippet in Python. It seems to be a function that preprocesses an image for object detection. The function takes an image as input and returns a preprocessed image and a list of bounding boxes.

Image Error *#16*

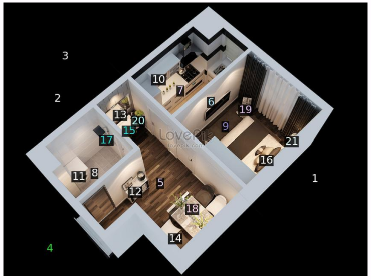


Text Prompt: What time is it on that watch?

LLaVA-NEXT-34B: The watch in the image is displaying a time that appears to be 10:11. The hour hand is pointing to the number 10, and the minute hand is pointing to the number 11.

Table 13: Failure Cases.

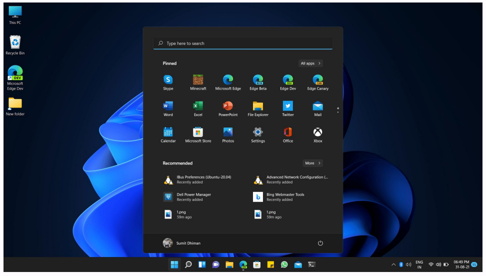
Image Error *#17*



Text Prompt: where to find food?

Gemini-Pro-Vision: The food is in the refrigerator, which is located in the kitchen. The kitchen is the room with the stove and oven.

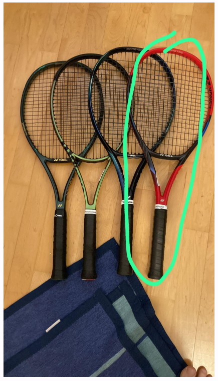
Image Error *#18*



Text Prompt: What time is it?

LLaVA-NEXT-34B: The time is 06:54 PM.

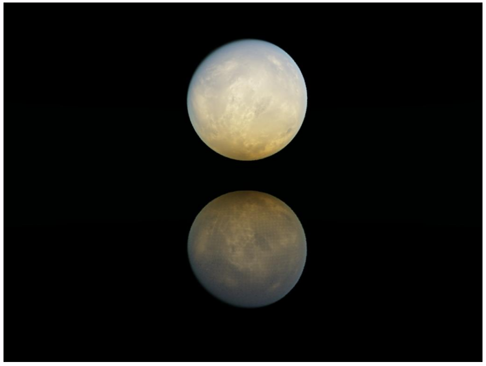
Image Error *#19*



Text Prompt: What Brand is highlighted

LLaVA-NEXT-7B: The brand highlighted in the image is Ḧead..

Image Error *#20*



Text Prompt: What causes this and which is the real one?

Qwen-VL-Chat: It is funny because the guy who is not in a white t-shirt is saying that he is in a white t-shirt.

D Data Analysis

Position and Length Biased Human Preferences Human preferences on models are known to be biased to the length of model responses. Previous work [12] show not strong correlation found between length and rank at model-wisely. To further understand each voting point in WildVision-Arena. We plot Figure 9 to show the winner distribution over token length. The Blue line represents the density of Model B wins over Model A, and the x axis be the token length difference which substract length of Model B from Model A. This plot help decouple the cofounders of position bias and length bias. And when the token length difference is more obvious, the preferences are slightly biased to length output model. And this effect is marginal when both model have long output.

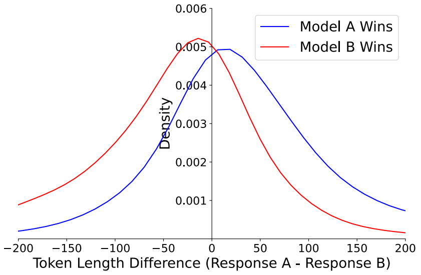


Figure 9: Winner Density Plot

Battles Trend In Figure 10, we plot the number of votes per day with a date cutoff at May 29, 2024. On average, WildVision-Arena got 71 votes from the users per day.

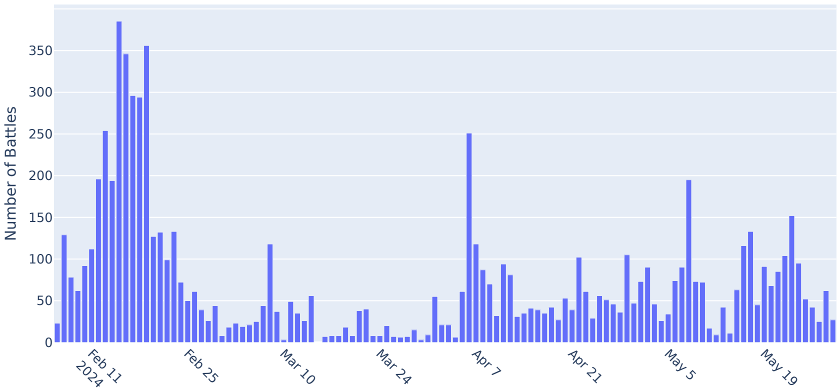


Figure 10: Number of Votes Per Day

Model Chats We visualize number of conversations per model in Figure 11.

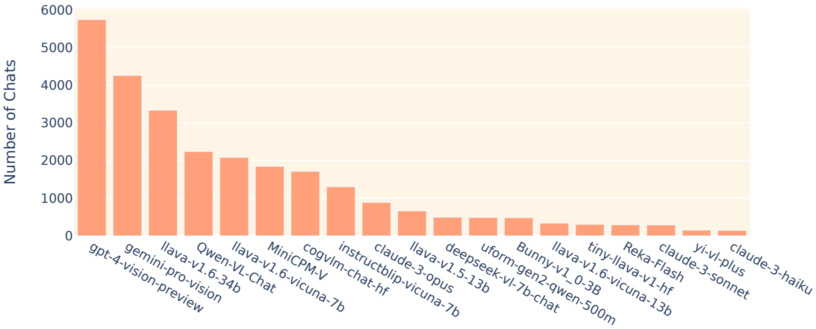


Figure 11: Number of Chats Per Model

E Prompt Template

E.1 Taxonomy Annotation

In Table E.1, we show the template fed into GPT-4v to annotate the question category and image domain for each data sample in WildVision-Arena as in Figure 2 and Figure 3.

[Image] <image>

[Question] What is the color of the main object in the image?

[System] Given the image and following text question, please classify the content according to the specified taxonomy:

Question Categories:
Descriptive - General Description (Provide a broad overview of what the image contains.) ...
Recognition - Object Recognition (What objects are present in the image?) ...
Instructive - How-to Guides (How do I obtain what's depicted in the image?) ...
Analytical - Data Analysis (Analyze the data presented in the image.) ...
Comprehensive - Cultural Analysis (Analyze the cultural significance of the image.) ...
Interactive - Bug Fixing (Fix the bug in the code depicted in the image.)
Creative - Music and Composition (Compose a song inspired by the image.)

Image Domains:
Urban - Cityscapes, Infrastructure, Public Spaces, Buildings, Transportation, Street Scenes
People - Portraits, Crowds, Faces, Selfies, Group Photos
Event - Cultural Events, Historical Events, Social Gatherings, Performances, Sports, Fashion, Lifestyle
Objects - Accessory, Vehicles, Sports Equipment, Kitchenware, Food, Furniture, Electronics, Appliances, Household Tools, Musical Instruments, Art Supplies, Office Supplies
Entertainment - Games, Movies and TV Shows, Media and Communication, Web and Mobile Apps Screenshots
Expert - Art and Design, Business, Science, Health and Medicine, Humanities and Social Science, Tech and Engineering

Please analyze the text and image provided and classify them into the appropriate category and subcategory, as well as the main image domain and subdomain, based on the taxonomy above. Please only reply with four values 1. question category, 2. question subcategory, 3. image domain, 4. image subdomain) in a string separated by [&]. For example, "Descriptive[&]Object Description[&]Natural[&]Landscapes".

[Output] Analytical[&]Attribute-based Question Answer[&]Objects[&]Furniture

E.2 VLM Voting

In Table E.2, we show the template used to generate the pairwise preference by utilizing GPT-4V as a local evaluator 3.2.

[Image] <image>

[Question] What is the color of the main object in the image?

[Model Assistant A's Response] Blue.

[Model Assistant B's Response] Red.

[System] Please act as an impartial judge and evaluate the quality of the responses provided by two model assistants to the user question displayed in [Question]. You should choose the assistant that follows the user's instructions and answers the user's questions better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any positional biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. Reply with "leftvote" if you find assistant A better, "rightvote" if assistant B is better, "bothbad_vote" if both responses are wrong, and "tievote" if both assistants provide equally satisfactory answers. If you are unable to make a decision, please reply with "NA".

[Evaluator Output] leftvote

E.3 WildVision-Bench Evaluator

In Table E.3, we show the template used to prompt judges to generate the pairwise preference by utilizing GPT-4o as a judge. We have defined have different judge results, which corresponds to the "Better+", "Better", "Tie", "Worse", and "Worse+" respectively in Table 4.

[System] Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below. You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is better.

Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.

When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information.

Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.

Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt.

After providing your explanation, you must output only one of the following choices as your final verdict with a label:

1. Assistant A is significantly better: [[A»B]]
2. Assistant A is slightly better: [[A>B]]
3. Tie, relatively the same: [[A=B]]
4. Assistant B is slightly better: [[B>A]]
5. Assistant B is significantly better: [[B»A]]

Example output: "My final verdict is tie: [[A=B]]".

[User] {question_1}

[Image] <image>

<|The Start of Assistant A's Answer|>
{answer_1}
<|The End of Assistant A's Answer|>

<|The Start of Assistant B's Answer|>
{answer_2}
<|The End of Assistant B's Answer|>

F Discussions

F.1 Limitations

Although our platform integrates a variety of multimodal models for convenient comparison, it inevitably omits some recently released models, potentially limiting the breadth of insights available. Additionally, the platform's stress testing is inadequate; scaling up is imperative to handle the increasing volume of user queries each day. There is also a critical need to balance the protection of third-party models with ensuring that model responses remain unbiased and true to their design. Despite logging data for research purposes and informing users accordingly, ongoing efforts are required to enhance system security to prevent data leaks.

F.2 Societal Impact

WildVision Arena serves as a dynamic benchmarking tool, embracing crowd-sourced input from a diverse range of users. However, biases persist, particularly among English-speaking users—a reflection of some models' linguistic limitations—and among those with a specific interest in multimodal research. Efforts are underway to refine the interface, aiming to broaden participation and reduce existing biases. By enhancing accessibility and user engagement, we strive to create a more inclusive platform that better represents global perspectives.

G Accessiblity of Datasets

G.1 Dataset Documentation and Intended Uses

To interact with models and submit votes, visit Hugging Face Vision Arena[1]. To view the live leaderboard, navigate to the leaderboard tab on the same page. Data can be accessed for downloading and viewing at WildVision Arena Data[2].

G.2 Maintenance Plan

The live leaderboard of WILDVISION-ARENA is updated every three hours. The data will be continually updated at WildVision on Hugging Face[3]. The code for the platform will be open-sourced at WildVision-Bench Github repo[4] and welcome community effort. The voting data and code for the evaluation will be provided to facilitate easy reproduction of the leaderboard.

G.3 Author Statement

We confirm that we bear all responsibility in case of violation of rights during the collection of data on WILDVISION-ARENA and WILDVISION-BENCH. We will take appropriate action when needed.

---

[1] https://huggingface.co/spaces/WildVision/vision-arena

[2] https://huggingface.co/datasets/WildVision/wildvision-arena-data

[3] https://huggingface.co/WildVision

[4] https://github.com/WildVision-AI