

# HRSScene: How Far Are VLMs from Effective High-Resolution Image Understanding?

## Supplementary Material

### 1. Dataset Details

**Autonomous Driving** We extract samples from the *MME-Realworld* dataset to evaluate a model’s embodied intelligence, focusing on perception tasks such as distant object perception, attribute recognition, and counting, as well as reasoning tasks including intention prediction, interaction relation understanding, and driver attention prediction.

**Monitoring** Extracted from *MME-Realworld*, this dataset features images taken from public safety cameras in diverse scenarios. It features realworld challenges including varying object scales and partially out-of-view objects captured from different view points across day and night.

**Document Parsing** For text recognition in images, we adopt *DocStruct4M*, which focuses on structure-aware parsing of complex document data in images across five domains: documents, webpages, tables, charts, and natural images.

**Fine-grained Perception** We select *HR-Bench* for fine-grained perception in high-resolution images. It poses single-instance and cross-instance perception tasks. The dataset is available in two resolution versions (4K and 8K), with the 8K images cropped around the relevant objects to produce the 4K versions. We select samples from both versions.

**Aerial Images** *HRVQA* is selected for aerial image understanding, it features images of a 1K spatial resolution and QA pairs that span 10 question types (Number, Yes/No, etc.) and 27 category concepts (Vehicles, Urban area, Water bodies, etc.).

**Image Quality** To evaluate models on quality assessment of daily-life pictures, we select *HRIQ* designed for Blind Image Quality Assessment (BIQA) based on human perceivable factors like blur, exposure, noise, etc. The label is a human aligned Mean Opinion Score (MOS) on a scale of 0 to 5 given as options. We also design a custom prompt to instruct the model about the task and the response format.

**Infographics** Infographics contain a mix of textual and visual elements arranged in complex layouts. We leverage samples from *InfographicVQA* to test a model’s ability to recognize and jointly reason over multiple spans of information present in infographics.

**Tissue Diagnosis** Automatic analysis of tissue samples can accelerate clinical diagnosis and treatment. To do this, we extract samples from *LungHist700*, a collection of histopathological lung tissue images for the classification of lung malignancies. We design a custom prompt to instruct the model on the task, the options (seven classes), and the

response format.

**Multi-Image** We choose *MuirBench* for its diverse tasks and multi-image relationships. To enable a single high-resolution image input, we combine multiple images in each sample into a grid on a canvas. In addition, we select only samples with answers and remove any unanswerable questions.

**Chart Comprehension** Applying Large Multimodal Models (LMMs) to charts enables efficient information processing and extraction of insights. Although we have collected chart data from other datasets, we select *NovaChart* for its comprehensiveness, featuring 18 different chart types and 15 chart-related tasks.

**Visual Difference** Describing differences between image sets is crucial in many real-world applications (cite). We repurpose *VisDiffBench* by selecting smaller subsets of 20 samples from the original image sets and creating a high-resolution image as a 4x10 grid, with the first two rows occupied by images from set 1 and the last two rows by images from set 2.

**Medical Image** A VQA dataset for radiology images of various types (X-rays and CT scans) covering the chest and abdominal regions with diverse question about size, modality, abnormality, etc.

**Telescope Image** The *Galaxy* dataset we use contains the images captured by a bubble telescope. We annotate this dataset from scratch with a question and four options.

**CAD** Contains floor-plan drawings of various architecture projects including residential buildings, schools, hospitals, and offices. It shows high variance in style and appearance of objects or symbols.

**PANDA** This dataset features high-resolution images with a wide Field-of-View (FoV) in outdoor scenarios, capturing pedestrians with varying crowd densities, poses, trajectories, and occlusions.

**V\*** A dataset for testing models on perceiving small details in High-Resolution images of real-life scenarios. Sub-tasks include attribute identification and spatial relationship reasoning of small very small objects.

**MileBench** We extract samples from MileBench, which evaluates multi-modal long-context understanding involving multiple images. The model must retain and integrate contextual information from extended inputs to answer questions accurately. The subtasks feature images that are temporally or semantically related.

**OCR in the Wild** Text recognition in real-world outdoor scenarios, such as streets and shops, involving the percep-

tion of advertisements, signage, identity information, and other textual elements. The samples are extracted from *MME-Realworld*.

**Remote Sensing** Extracted from *MME-Realworld*, this tests perception in high-resolution images with rich details, encompassing object counting, color recognition, and spatial relationship understanding.

**Chart and Diagram** Unlike other chart datasets, this dataset presents highly complex chart data, such as financial reports, which feature extensive numerical information and mathematical content. It evaluates both perception and reasoning capabilities of models. The perception tasks involve locating values in diagrams and tables, while the reasoning tasks include identifying maximum and minimum values, performing calculations, and predicting trends. The samples are extracted from *MME-Realworld*.

**ArtBench** Contains artwork from 10 different artistic styles: Baroque, Surrealism, Post Impressionism, Realism, Romanticism, Impressionism, Art Nouveau, Expressionism, Renaissance, and Ukiyo-e. The correct artistic style along with few other distractor choices are given options.

**Museum** To assess models on art understanding, the *MAME* dataset comprises of various artworks and their corresponding medium (the various materials and techniques used to create the artwork). The dataset exhibits high intra-class variance, requiring models to pay close attention to fine-grained details.

**Animals** This dataset presents the task of counting various types of waterfowl using high-resolution aerial images of water bodies. This task is relevant for surveying waterfowl and reduces the manual effort.

**Product Anomaly Detection** Evaluating LMMs on their ability to identify defective (anomalous) products presents a highly industry-relevant task. This dataset not only supports anomaly detection but also includes additional subtasks for anomaly analysis, such as defect type classification, defect localization, and severity assessment.

**Grass** Automated inspection of vegetation, such as signal grass (*Urochloa*), is crucial for farmers and promotes sustainable agriculture. To assess models in this real-world application, we adopt the task of phenological stage classification and raceme counting in high-resolution RGB images of *Urochloa*.

**Diagnosis Datasets** For WhiteBackground dataset, we first pick 500 samples from VQAv2 dataset. Then, we combine each sample with white background images of different sizes. In this paper, we include 1x1 (no white background), 3x3, 5x5, 7x7, and 10x10 versions.  $N \times N$  indicates the needle image is combined with  $N \times N - 1$  white background images of the same size to form the entire image. In this case, the needle image has  $N \times N$  positions for each sample. We run experiments to observe the difference in performance in each position and measure Regional Diver-

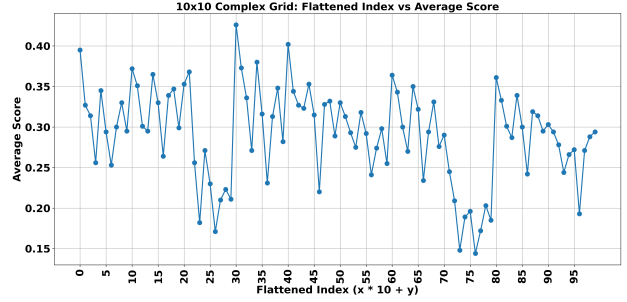


Figure 1. The performance of all models with an increase of patch id. Unlike lost-in-the-middle, no significant pattern can be observed.

gence. Similarly, in ComplexGrid, we use similar images to fill the background rather than white background images. To pick out the most similar images, we use BLIP [13] to rank the similarity between the needle image and all images in the validation set of VQAv2. And use the most similar  $N \times N - 1$  images as the haystack.

## 2. Ablation on Lost-in-the-Middle

To test whether the observed U-shape is a trivial extension of existing work [15], we further evaluate the model with flattened distance, the metric used in the original lost-in-the-middle that measures the linear distance of the starting token and needle tokens in the input. Since VLMs use a vision transformer [18] that inputs the image as linear patches, similarly, we measure the distance by counting how many patches the needle is from the first patch.

The results are shown in Figure 1. As shown, no significant pattern can be observed with the increasing of the patch distance, showing that the proposed phenomenon is not the same as the original lost-in-the-middle.

## 3. Comparison between GPT and Qwen

As shown in Table 2, we find that GPT 4o outperforms Qwen2-VL 72B on HRVQA with 1k resolution, while Qwen2-VL performs much better on Galaxy with images as large as 20k resolution. Qwen can input images with native resolution, while GPT has a size limit of 5 MB. This shows that due to the native resolution support of Qwen, it obtains SOTA, even general capability might not be the best.

## 4. Model Details

### 4.1. Implementation Details

We include a total of 28 VLMs in our experiment. The details are in supplementary materials. This includes one **Phi-3.5** [1], two **DeepSeek** [4], seven **InterVL2** [5–7, 10], two **Qwen2-VL** [18], two **MolMo** [8], one

Table 1. Overview of 25 real-world datasets and their statistics. \* indicates that the dataset is reannotated.

Dataset Name	Explanation	Capability	# Samples	Min Res	High Res	Avg. Res
Autonomous_Driving	Street View	Small Object Understanding	300	5760x1200	5760x1200	5760x1200
DocStruct4M*	Text document	OCR	296	1024x1024	4000x28990	1733x2675
HR-Bench	Daily photos	Small Object Understanding	397	4032x1152	7680x7680	5740x4458
HRVQA*	Aerial Image	Spactial relation, Small Objects	273	1024x1024	1024x1024	1024x1024
HRIQ	Long range picture	Small Object Understanding	300	2880x2160	2880x2160	2880x2160
InfographicVQA*	Graophic layout document	OCR	300	1024x1024	6250x9375	1970x3881
LungHist700	Microscope Medical Image	Domain Knowledge	308	1600x1200	1600x1200	1600x1200
MuirBench*	Multi-imge combination	Multi-image Reasoning	300	1064x1204	16062x7704	3072x2334
NovaChart*	Chart Image	Chart Understanding	297	2000x1600	3000x2147	2006x1600
Video_Monitoring	Street monitor	Small Object Understanding	300	1280x1024	2048x2048	1989x1460
VisDiffBench*	Multi-image combination	Multi-image Reasoning	150	5220x1648	5220x2088	5220x2085
VQA-RAD	Medical x-ray image	Domain Knowledge	225	1024x1024	2321x1384	1041x1230
Galaxy*	Telescope Image	Counting	87	1435x732	29566x14321	4828x4078
OCR_in_the_Wild	Street brands	Small Object OCR	300	1056x1056	7680x4320	2282x1867
Remote_Sensing	Shop signs	Small Object Understanding	300	1272x1419	11500x7500	5788x4536
Diagram_and_Table	Chart inside large image	Small Chart Object Understanding	300	1201x1086	2481x3507	2337x1521
VStar_Bench	Daily photos	Image Search	232	1080x1439	7500x5000	2357x1683
MAME	Museum artwork	Domain Knowledge	300	1109x1043	15649x8900	3124x3200
Izembek	Remote sensing of Zoo	Counting	300	8688x5792	8688x5792	8688x5792
ArtBench	Scanned Painting	Domain Knowledge	306	1083x1024	9449x6496	1982x2017
Grass	Argiculture Image	Counting	300	4224x3168	4224x3168	4224x3168
MMAD	Daily photo	Reasoning	300	1024x1024	3024x3024	1918x1777
MileBench	Video frame	Image Reseasoning	300	1600x800	6400x6400	3096x2506
PANDA	Public Monitor for Crowd	Crowd Counting	300	24853x13983	35503x26627	27002x16152
CAD*	Interior Design	Spactial relation, Counting	297	2000x2000	2000x2000	2000x2000
Total	N/A	N/A	7068	1024x1024	35503x26627	5359x5395

Table 2. Fine-grained comparison between Qwen 72B and GPT 4o on two datasets.

	HRVQA (1k)	Galaxy (20k)
Qwen 72B	69.59	<b>80.80</b>
GPT 4o	<b>73.82</b>	68.68

**LLaVA-Onevision** [14], one **LLaVA-HR** [16], four **Llava-Next** [20], two **Llama-3.2** [9], and two **GPT-4o** [12], two **Gemini** [17], and two **Claude** [3].

Specifically, **Phi-3.5** [1] is a lightweight model designed for efficient language understanding and generation. We include **Phi 3.5 vision instruct** [1] for experiments. **DeepSeek Janus Pro 7B** [4] is a model that integrates multi-modal reasoning capabilities. **DeepSeek-VL2** [19] is a vision-language model, with **deepseek vl2 27B** included in our evaluation. **InterVL2** [5–7, 10] is a family of multi-modal models ranging from small to large-scale by OpenGVLab. We include **InterVL2 1B**, **InterVL2 2B**, **InterVL2 4B**, **InterVL2 8B**, **InterVL2 26B**, **InterVL2 40B**, and **InterVL2 Llama3 76B** for experiments. **Qwen2-VL** [18] is a vision-language model, and we consider both **Qwen2 VL 7B Instruct** and **Qwen2 VL 72B Instruct**. **MolMo** [8] is a series of models designed for molecular and scientific applications. We include **Molmo 72B 0924** and its distilled variant, **Molmo 7B D 0924**. **LLaVA-**

**Onevision** [14] is an open-source multimodal LLM, we selected **llava-onevision-qwen2-72b-ov-hf** model for our experiments. **Llava-Next** [20] is an evolution of LLaVA, and we include **llama3-llava-next-8b-hf**, **llava-v1.6-vicuna-13b-hf**, **llava-v1.6-34b-hf**, and **llava-next-72b-hf** in our experiments. **Llama3.2** builds on the Llama architecture with enhanced scalability. We include **Llama-3.2-11B-Vision-Instruct** and **Llama-3.2-90B-Vision-Instruct** in our experiments. **GPT** [2] includes versions optimized for both efficiency and performance, with **GPT 4o** and **GPT 4o-mini** selected. Gemini is a family of LLMs, and we evaluate **Gemini 2.0 Flash** and **Gemini 1.5 Pro** [17]. **Claude** is a family of LLMs known for its strong reasoning and safety features. We include two models in ascending order of capability: **Claude-3-haiku** and **Claude-3.5-sonnet**.

For the evaluation, for WhiteBackground, we follow the accuracy in VQAv2[11]. For ComplexGrid dataset, we prompt the model to generate the column and row of the needle image and compare with the gold column and row using an exact match. For real-world datasets, since they are MCQ-based, we directly use exact math as metrics.

## 5. Performance Details on Real-world Datasets

Table 3 and Table 4 display the performance of all VLMs on every real-world dataset. The scores are the average performance of all samples in val, test, testmini splits.

Table 3. Performance of all VLMs on every real-world dataset (Part 1).

	Drive	DocStr	HR-B	HRVQA	HRIQ	InfoQ	Lung	Muir	Nova	Monitor	VisDiff	RAD
Random	20.00	25.02	25.00	25.06	20.00	25.06	14.29	23.38	23.70	20.00	25.00	33.33
Calude3 Haiku	27.35	62.05	29.43	62.32	29.15	56.14	14.11	50.00	57.21	19.40	84.09	46.20
Calude3.5 Sonnet	25.21	79.46	48.42	69.57	36.77	83.33	23.24	65.68	81.22	28.02	92.05	63.92
Gemini1.5 Pro	29.49	63.39	59.18	73.91	39.01	62.28	32.37	57.20	73.80	29.74	75.00	65.19
Gemini2.0 Flash	38.03	67.41	64.56	74.88	43.05	88.16	46.89	63.56	68.12	34.91	82.95	58.86
DeepSeek-VL2 27B	37.18	58.48	61.08	54.59	34.53	65.79	15.35	61.02	66.81	43.10	87.50	63.92
GPT-4o	32.05	67.86	55.70	72.95	44.39	74.12	1.24	58.47	73.80	34.91	89.77	51.90
GPT-4o mini	30.77	54.91	47.78	74.40	46.19	61.40	26.56	42.80	64.63	25.86	65.91	30.38
InternVL2 1B	22.22	35.27	38.92	36.23	21.97	36.84	12.45	30.51	32.31	21.98	28.41	54.43
InternVL2 2B	38.89	42.41	42.41	64.73	34.08	51.32	12.45	30.51	54.15	26.29	23.86	60.13
InternVL2 4B	35.90	56.25	46.84	48.79	29.15	63.60	19.09	58.90	58.95	35.78	85.23	73.42
InternVL2 8B	37.61	67.86	49.05	60.87	39.46	71.49	18.67	52.54	58.08	28.02	89.77	70.25
InternVL2 26B	42.74	68.30	60.44	58.45	32.74	70.61	17.01	58.90	62.45	40.09	85.23	67.09
InternVL2 40B	37.61	72.77	66.14	67.63	36.32	84.21	13.69	64.83	68.56	39.22	95.45	71.52
InternVL2 76B	38.46	69.64	59.81	58.45	39.46	84.65	14.11	65.68	55.90	41.81	94.32	70.89
DeepSeek-Janus 7B	30.34	39.73	31.96	51.69	27.35	41.67	14.94	48.31	41.92	22.41	51.14	64.56
Llama3.2 11B	32.91	58.93	52.85	70.05	21.52	67.11	25.73	46.19	63.32	30.60	81.82	63.92
Llama3.2 90B	31.62	72.77	54.11	74.88	23.32	71.05	17.01	54.66	69.00	34.48	82.95	71.52
Llava-HR 7B	31.20	28.57	41.46	55.07	28.70	37.72	14.52	25.85	38.86	31.90	47.73	44.94
Llava-Next 8B	34.19	41.96	44.62	67.15	21.97	47.81	12.45	40.25	43.23	28.45	79.55	60.76
Llava-Next 13B	28.63	48.21	40.82	54.59	31.39	46.05	13.28	45.76	55.46	37.50	80.68	68.35
Llava-Next 34B	29.91	66.96	53.80	66.18	36.32	59.65	15.77	58.90	65.50	31.90	80.68	65.82
Llava-Next 72B	29.91	67.41	51.58	63.77	34.08	63.16	14.94	46.61	64.63	35.78	85.23	65.19
Llava-OneVision 72B	32.91	72.32	62.34	68.12	46.64	80.70	15.35	66.10	69.87	33.19	75.00	78.48
Phi3.5 4B	32.91	54.02	45.89	65.22	43.50	62.28	7.05	46.19	58.08	31.03	89.77	67.72
MolMo 7B-D	33.76	52.68	46.52	55.07	34.98	68.86	13.69	43.64	54.59	32.33	64.77	55.06
MolMo 72B	33.33	69.64	56.01	71.01	32.29	78.51	14.52	63.56	70.74	40.95	84.09	65.19
Qwen2-VL 7B	35.04	85.27	68.35	75.36	39.46	87.28	14.11	67.37	72.93	37.07	94.32	84.18
Qwen2-VL 72B	32.48	68.30	62.34	69.08	47.98	71.49	15.35	63.98	66.38	34.48	94.32	74.68
Human	40.00	82.00	96.00	68.00	40.00	92.00	14.00	84.00	88.00	67.00	100.00	33.33

Table 4. Performance of all VLMs on every real-world dataset (Part 2).

	Galaxy	Remote	OCRW	D&T	VStar	MAME	Izem	ArtB	Grass	MMAD	Mile	PANDA	CAD
Random	25.00	20.00	20.00	20.00	34.09	10.00	22.33	11.11	20.00	29.75	27.67	25.00	25.03
Calude3 Haiku	74.07	10.13	51.11	45.02	28.66	74.11	20.60	56.03	16.24	61.90	51.54	12.83	43.44
Calude3.5 Sonnet	59.26	32.60	67.11	71.43	46.34	83.04	22.32	61.64	17.09	68.40	62.56	16.81	81.45
Gemini1.5 Pro	81.48	37.00	75.11	51.95	65.24	83.04	19.31	69.40	32.48	70.13	59.47	18.58	67.42
Gemini2.0 Flash	51.85	48.02	75.11	76.62	71.95	86.61	22.32	59.91	29.91	71.00	66.52	30.09	83.26
DeepSeek-VL2 27B	81.48	54.19	63.56	56.28	67.07	75.45	30.47	64.66	21.37	73.59	60.35	34.51	75.57
GPT-4o	85.19	33.92	76.89	51.52	49.39	82.14	27.90	65.95	20.94	72.73	59.47	23.01	59.28
GPT-4o mini	77.78	13.66	60.89	44.59	51.22	75.45	24.46	59.05	16.67	71.43	56.39	15.49	46.61
InternVL2 1B	44.44	14.98	39.56	25.11	32.93	31.25	40.34	22.41	20.94	51.95	49.34	18.58	39.82
InternVL2 2B	66.67	37.44	57.78	35.06	52.44	55.36	18.88	53.02	16.24	54.98	63.88	22.12	40.27
InternVL2 4B	59.26	39.21	60.89	52.81	47.56	66.07	21.03	55.17	28.63	65.37	63.88	30.53	61.09
InternVL2 8B	70.37	34.80	68.89	35.50	64.02	66.52	18.88	56.90	32.05	66.67	64.76	18.14	60.63
InternVL2 26B	77.78	46.26	74.67	57.14	68.29	78.13	24.46	56.47	23.93	71.86	69.16	26.11	66.52
InternVL2 40B	74.07	50.66	77.78	58.01	75.00	82.14	16.31	65.09	34.19	74.89	72.69	23.01	76.02
InternVL2 76B	70.37	49.78	74.22	58.01	73.17	81.70	23.61	64.22	17.52	77.49	71.37	10.62	63.35
DeepSeek-Janus 7B	77.78	37.89	50.67	20.78	42.68	66.52	19.31	49.57	25.64	67.10	44.05	34.51	45.25
Llama3.2 11B	70.37	45.82	66.22	48.48	56.10	72.32	18.45	58.62	34.62	67.97	60.79	23.01	66.06
Llama3.2 90B	74.07	38.33	72.00	44.59	60.37	82.59	24.03	59.48	19.23	71.00	62.56	19.03	70.14
Llava-HR 7B	48.15	22.47	44.00	20.78	39.02	50.00	30.90	39.22	20.94	54.98	41.85	15.49	29.41
Llava-Next 8B	70.37	44.93	52.44	26.41	59.76	60.71	30.47	40.95	16.24	67.53	56.83	18.14	40.72
Llava-Next 13B	77.78	31.28	56.44	35.50	49.39	54.46	18.88	43.97	16.67	64.94	49.78	12.39	30.77
Llava-Next 34B	88.89	43.17	59.11	37.23	59.15	74.55	20.17	56.03	40.60	74.46	57.27	21.24	63.80
Llava-Next 72B	74.07	46.70	58.67	33.77	57.93	73.21	17.60	59.91	19.23	74.89	58.59	14.60	49.77
Llava-OneVision 72B	88.89	44.49	73.33	52.81	78.05	80.36	27.90	62.93	41.45	75.32	68.28	21.68	61.54
Phi3.5 4B	81.48	40.97	59.11	47.19	53.05	59.38	9.87	56.03	29.06	69.26	53.30	19.91	56.56
MolMo 7B-D	55.56	49.34	64.00	33.77	71.34	51.34	24.89	41.81	24.79	72.29	56.39	22.57	50.68
MolMo 72B	85.19	51.98	75.56	35.93	71.95	64.73	35.19	53.88	47.01	76.19	56.83	26.11	62.44
Qwen2-VL 7B	85.19	50.66	78.22	67.10	84.15	83.93	33.48	62.50	26.07	76.19	72.69	28.32	81.45
Qwen2-VL 72B	70.37	43.61	78.67	52.38	76.22	82.59	40.77	61.21	26.92	73.59	67.40	15.04	64.71
Human	54.00	87.00	68.00	93.00	100	76.00	20.00	57.00	43.00	75.00	86.00	46.00	93.00

## 6. Prompts and Metrics

For ComplexGrid dataset, our prompt is “The image is composed of multiple sub-images. The left upper corner is row 1 column 1. We also add the row and column numbers under each image. You need to identify the sub-image that best suits the caption: {caption}, returning the row and column id of the needle sub-image in this format: <row>ROW</row><col>COL</col>, such as <row>3</row><col>2</col>”. We ask the model to answer with the HTML tag because we could use BeautifulSoup to parse the tag to get a clean prediction to avoid evaluation bias. For the real-world dataset, we also adopt the same idea as tag parse. Our prompt is “question n Give an answer with this format: <ans>ANSWER</ans>, no redundant words. For example: <ans>A</ans>”. We use exact math as our metrics during the evaluation.

## 7. Examples

Table 5 to 27 show examples from HRScene real-world datasets. We compress the images to display them in the paper.

Table 5. Example from HRScene – ArtBench



The painting in the picture belongs to which of the following categories?

- (A) Surrealism
- (B) Expressionism
- (C) Realism
- (D) Romanticism
- (E) Art Nouveau
- (F) Ukiyo E
- (G) Post Impressionism
- (H) Impressionism
- (I) Baroque

Answer: H



The painting in the picture belongs to which of the following categories?

- (A) Ukiyo E
- (B) Art Nouveau
- (C) Post Impressionism
- (D) Realism
- (E) Impressionism
- (F) Baroque
- (G) Romanticism
- (H) Expressionism
- (I) Surrealism

Answer: E

Table 6. Example from HRScene – Autonomous Driving



What is motion of the pedestrian wearing blue top on the left?

- (A) crossing the crosswalk
- (B) standing
- (C) jaywalking (illegally crossing not at pedestrian crossing)
- (D) walking on the sidewalk
- (E) The image does not feature the object

Answer: B

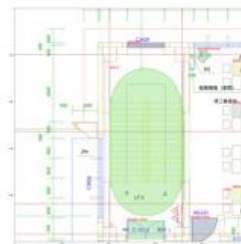


What is motion of the purple sedan on the right?

- (A) parked
- (B) moving
- (C) stopped
- (D) other
- (E) The image does not feature the object

Answer: E

Table 7. Example from HRScene – CAD



How many doors are there in the image?

- (A) 1
- (B) 0
- (C) 2
- (D) 3

Answer: A



What is the shape of the shadow at upper left corner of the image?

- (A) L-shape
- (B) Oval
- (C) Circle
- (D) Square

Answer: A

Table 8. Example from HRScene – Diagram and Table

What's the data of Shipping Costs of 2028 Year 5 in the table Profit per kg NH3 Analysis?

- (A) -0.51
- (B) -0.52
- (C) -0.53
- (D) -0.54
- (E) This image doesn't feature the data.

Answer: D


What is the revenue of Pigs Feed in year 5 in the Revenue Sources table?

- (A) 4.548,625
- (B) 4.223.063
- (C) 3.710.817
- (D) 4.058.442
- (E) The image does not feature the number.

Answer: D




Table 9. Example from HRScene – DocStruct4M



Read the following text: <doc>CALL FOR NOMINATIONS  
BILINGUAL INSTRUCTIONAL ASSISTANT OF THE YEAR AWARD  
[omitted]  
Which of the following options is correct?  
(A) the nominee's outstanding [omitted]  
14 </doc>  
(B) the nominee's outstanding [omitted]  
14 </doc>  
(C) the nominee's outstanding [omitted]  
14 </doc>  
(D) the nominee's outstanding [omitted]  
14 </doc>


Answer: D



Which of the following sentences is present in the image?  
Which of the following options is correct?  
(A) <ocr>CONTACTS </ocr>  
(B) <ocr>CONTACT </ocr>  
(C) <ocr>CONTRACT </ocr>  
(D) <ocr>CONVENANT </ocr>


Answer: C

Table 11. Example from HRScene – Grass



Based on the plant in the image, which growth stage does it belong to, and how many racemes does it have?  
(A) Reproductive stage, more than 200  
(B) Reproductive stage, 10-100 (include 100)  
(C) Reproductive stage, 0-10 (include 10)  
(D) Reproductive stage, 100-200 (include 200)  
(E) Vegetative stage, no racemes

Answer: A




Based on the plant in the image, which growth stage does it belong to, and how many racemes does it have?  
(A) Reproductive stage, 10-100 (include 100)  
(B) Reproductive stage, more than 200  
(C) Reproductive stage, 0-10 (include 10)  
(D) Vegetative stage, no racemes  
(E) Reproductive stage, 100-200 (include 200)

Answer: B


Table 12. Example from HRScene – HR-Bench

Table 10. Example from HRScene – Galaxy




What type of celestial object is shown in the image? Please note that only clearly visible or distinguishable celestial bodies are counted.  
(A) Elliptical  
(B) star  
(C) Spiral  
(D) irregular

Answer: B




Does the galaxy have a distinct central core? Please note that only clearly visible or distinguishable celestial bodies are counted.  
(A) No  
(B) I don't know  
(C) Yes  
(D) two

Answer: C



What is the number displayed above the entrance where the woman is standing?  
(A) 27E  
(B) 37B  
(C) 27D  
(D) 27B

Answer: D



What is the color of the mailbox?  
(A) Green  
(B) Black  
(C) Red  
(D) Blue

Answer: D

Table 13. Example from HRScene – HRIQ



Assess the quality of a given image and predict a score that reflects the mean subjective human judgment of image quality. Some factors you may consider are distortions, such as Noise, Out-of-focus blur, Motion blur, Overexposure / Underexposure, Low contrast, Incorrect saturation, Sensor noise, and any combination of these distortions. Do not rely on meta-data or external references - your judgment should be based purely on visual quality.

- (A) 1 bad
- (B) 2 poor
- (C) 3 fair
- (D) 4 good
- (E) 5 excellent

Answer: D



Assess the quality of a given image and predict a score that reflects the mean subjective human judgment of image quality. Some factors you may consider are distortions, such as Noise, Out-of-focus blur, Motion blur, Overexposure / Underexposure, Low contrast, Incorrect saturation, Sensor noise, and any combination of these distortions. Do not rely on meta-data or external references - your judgment should be based purely on visual quality.

- (A) 1 bad
- (B) 2 poor
- (C) 3 fair
- (D) 4 good
- (E) 5 excellent

Answer: D

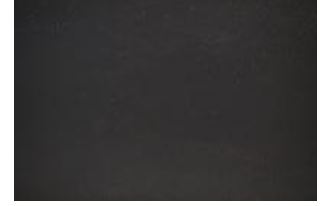
Table 15. Example from HRScene – Izembek



How many goose or other animals do you see in the image?

- (A) more than 400
- (B) 100-200
- (C) 200-300
- (D) 300-400

Answer: A

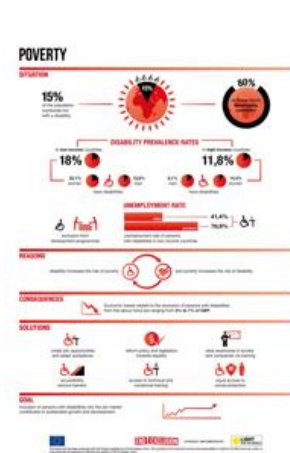


How many goose or other animals do you see in the image?

- (A) more than 400
- (B) 300-400
- (C) 200-300
- (D) 100-200

Answer: C

Table 14. Example from HRScene – InfographicVQA



what percent of people live without disability around the world according to the data given?

- (A) '80', '80%'
- (B) '79.9', '79.9%'
- (C) '15', '15%'
- (D) '85', '85%'

Answer: D

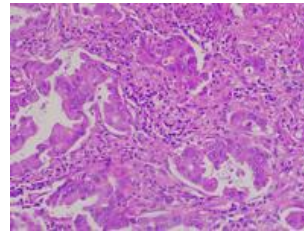


Which of these animals are shown in the image?

- (A) 'cow, fish'
- (B) 'cow, human'
- (C) 'cat, cow'
- (D) 'plane, apple'

Answer: A

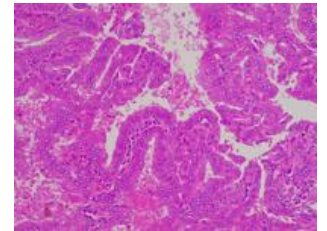
Table 16. Example from HRScene – LungHist700



Given the following histopathological image of lung tissue, classify the malignancy (if any) into one of the seven categories:

- (A) Normal tissue
- (B) Adenocarcinoma (Well-differentiated)
- (C) Adenocarcinoma (Moderately differentiated)
- (D) Adenocarcinoma (Poorly differentiated)
- (E) Squamous cell carcinoma (Well-differentiated)
- (F) Squamous cell carcinoma (Moderately differentiated)
- (G) Squamous cell carcinoma (Poorly differentiated)

Answer: B



Given the following histopathological image of lung tissue, classify the malignancy (if any) into one of the seven categories:

- (A) Normal tissue
- (B) Adenocarcinoma (Well-differentiated)
- (C) Adenocarcinoma (Moderately differentiated)
- (D) Adenocarcinoma (Poorly differentiated)
- (E) Squamous cell carcinoma (Well-differentiated)
- (F) Squamous cell carcinoma (Moderately differentiated)
- (G) Squamous cell carcinoma (Poorly differentiated)

Answer: B



Table 17. Example from HRScene – MAME



The artwork in the picture belongs to which of the following medium categories?

- (A) Hand-colored etching
- (B) Lithograph
- (C) Faience
- (D) Silk and metal thread
- (E) Graphite
- (F) Etching
- (G) Clay
- (H) Ivory
- (I) Woodcut
- (J) Oil on canvas

Answer: J



The artwork in the picture belongs to which of the following medium categories?

- (A) Lithograph
- (B) Oil on canvas
- (C) Ivory
- (D) Porcelain
- (E) Silver
- (F) Woodblock
- (G) Steel
- (H) Limestone
- (I) Marble
- (J) Iron

Answer: B

Table 18. Example from HRScene – MMAD



There is a defect in the object. Where is the defect?

- (A) On the top of the can
- (B) On the bottom of the can
- (C) Around the center region of the can, on the image of the potato chip
- (D) On the side of the can

Answer: C



There is a defect in the object. What is the appearance of the defect?

- (A) The defective capsule has a distinct non-conforming orange color.
- (B) The defective capsule has a shiny, translucent quality.
- (C) The defective capsule has a misshapen appearance.
- (D) The defective capsule has visible bubbles.

Answer: A

Table 19. Example from HRScene – MuirBench



What type of clothing was the man primarily seen wearing? <image1><image2><image3><image4><image5><image6><image7><image8>

- (A) None of the choices provided
- (B) Green and white jacket
- (C) Robe and shawl
- (D) Sweater

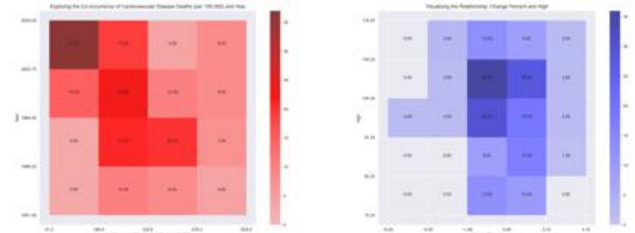
Answer: C

<image1>Which of the following images shares the same scene with the given image but contains the object dining table?

- (A) <image2>
- (B) <image3>
- (C) <image4>
- (D) None of the choices provided
- (E) <image5>

Answer: C

Table 20. Example from HRScene – NovaChart



Can you discern the type of chart used in this visualization? From the provided alternatives, please select the correct choice for the question above:

- (A) bivariate histogram
- (B) single-class scatter plot
- (C) radar chart
- (D) pie chart
- (E) univariate histogram

Answer: A

Can you provide the histogram value for the bin corresponding to the range  $x=[-4.0, -1.96]$  and  $y=[73.235, 82.2385]$ ?

- (A) 13
- (B) 9
- (C) 2
- (D) 3
- (E) 0

Answer: E

Table 21. Example from HRScene – OCR in the Wild



What is the content on the plaque in the center of the picture?

- (A) 安らぎの榎
- (B) 安らぎの庭
- (C) 安らぎの榎
- (D) 安らぎの庭
- (E) This image doesn't feature the content.

Answer: D



How long is this film in the picture?

- (A) 5.1
- (B) 2013
- (C) 148 min.
- (D) 143 min.
- (E) The image does not feature the content.

Answer: D

Table 22. Example from HRScene – PANDA



How many riding person(s) are in the image?

- (A) 35
- (B) 27
- (C) 23
- (D) 44

Answer: C



How many riding person(s) are in the image?

- (A) 11
- (B) 12
- (C) 21
- (D) 15

Answer: B

Table 23. Example from HRScene – Remote Sensing



What color is the second ship from top to bottom on the far right side of the picture?

- (A) White
- (B) Red
- (C) Green
- (D) Yellow
- (E) This image doesn't feature the color.

Answer: A



How many red cars are there in the parking lot in the middle of the bottom of the picture?

- (A) 1
- (B) 2
- (C) 3
- (D) 4
- (E) This image doesn't feature the count.

Answer: D

Table 24. Example from HRScene – VQA-RAD



Is the trachea midline?

- (A) Yes
- (B) No
- (C) Not specified

Answer: A



Is there evidence of an aortic aneurysm?

- (A) Yes
- (B) No
- (C) Not specified

Answer: B

Table 25. Example from HRScene – VStar Bench



What is the color of the car?

- (A) The color of the car is silver.
- (B) The color of the car is black.
- (C) The color of the car is red.
- (D) The color of the car is blue.

Answer: A



Is the flag blue and yellow or red and yellow?

- (A) The color of the flag is red and yellow.
- (B) The color of the flag is blue and yellow.

Answer: B

Table 26. Example from HRScene – Video Monitoring



What is the number of people in the image?(If a human maintains standing pose or walking, please classify it as pedestrian, otherwise, it is classified as a people.)

- (A) 97
- (B) 88
- (C) 52
- (D) 100
- (E) The image does not feature the people

Answer: E



What is the number of tricycles in the image?

- (A) 51
- (B) 97
- (C) 55
- (D) 74
- (E) The image does not feature the tricycles

Answer: E

Table 27. Example from HRScene – VisDiffBench



What is the difference between the first two rows of images and the last two rows?

- (A) Animal species (Dogs vs Cats)
- (B) Animal species (Cows vs Cats)
- (C) Background Colors (Green vs Blue)
- (D) Number of Objects (2 vs 3)

Answer: A



What is the difference between the first two rows of images and the last two rows?

- (A) Activity (Basketball vs Swimming)
- (B) Number of animals (1 vs 2)
- (C) Animal species (Cows vs Cats)
- (D) Activity (Soccer vs Swimming)

Answer: D

## References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL <https://arxiv.org/abs/2404.14219>, 2024. 2, 3
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [3] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024. 3
- [4] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 2, 3
- [5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 2, 3
- [6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [7] 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*, pages 24185–24198, 2024. 2, 3
- [8] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 2, 3
- [9] 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. 3
- [10] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024. 2, 3
- [11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3
- [12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3
- [13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2
- [14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3
- [15] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024. 2
- [16] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiewu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In *International Conference on Learning Representations (ICLR)*, 2025. 3
- [17] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 3
- [18] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3
- [19] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 3
- [20] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 3