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

Yusen Zhang, Wenliang Zheng, Aashrith Madasu, Peng Shi, Ryo Kamoi, Hao Zhou,
Zhuoyang Zou, Shu Zhao, Sarkar Snigdha Sarathi Das, Vipul Gupta, Xiaoxin Lu, Nan Zhang,
Ranran Haoran Zhang, Avitej Iyer, Renze Lou, Wenpeng Yin, Rui Zhang
Penn State University
{yfz5488, rmz5227}@psu.edu

https://yszh8.github.io/hrscene/

Abstract

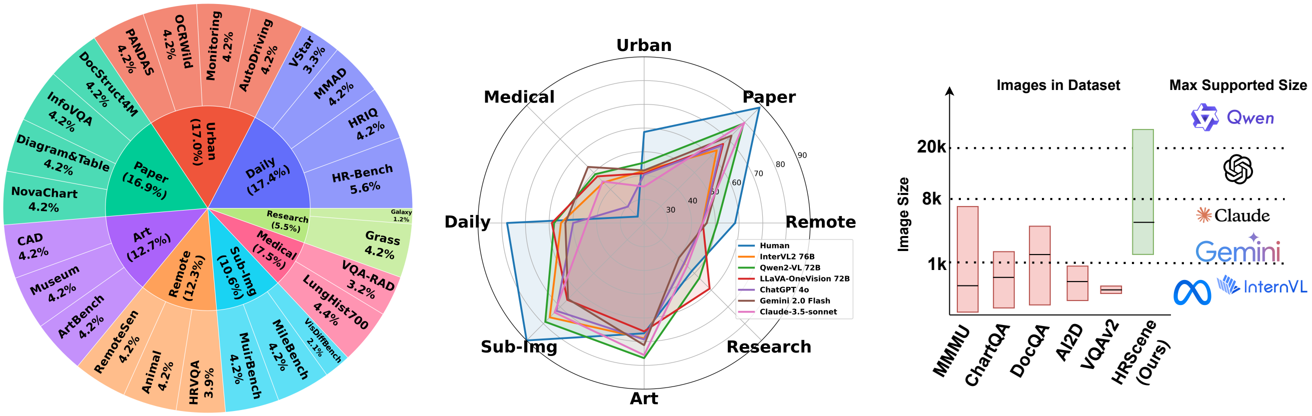
High-resolution image (HRI) understanding aims to process images with a large number of pixels, such as pathological images and agricultural aerial images, both of which can exceed 1 million pixels. Vision Large Language Models (VLMs) can allegedly handle HRIs, however, there is a lack of a comprehensive benchmark for VLMs to evaluate HRI understanding. To address this gap, we introduce HRScene, a novel unified benchmark for HRI understanding with rich scenes. HRScene incorporates 25 real-world datasets and 2 synthetic diagnostic datasets with resolutions ranging from 1,024 × 1,024 to 35,503 × 26,627. HRScene is collected and re-annotated by 10 graduate-level annotators, covering 25 scenarios, ranging from microscopic to radiology images, street views, long-range pictures, and telescope images. It includes HRIs of real-world objects, scanned documents, and composite multi-image. The two diagnostic evaluation datasets are synthesized by combining the target image with the gold answer and distracting images in different orders, assessing how well models utilize regions in HRI. We conduct extensive experiments involving 28 VLMs, including Gemini 2.0 Flash and GPT-4o. Experiments on HRScene show that current VLMs achieve an average accuracy of around 50% on real-world tasks, revealing significant gaps in HRI understanding. Results on synthetic datasets reveal that VLMs struggle to effectively utilize HRI regions, showing significant Regional Divergence and lost-in-middle, shedding light on future research.

1. Introduction

High-resolution image (HRI) understanding aims to process images with a large number of pixels [7]. It plays an important role in numerous scenarios, such as pathology [16], autonomous driving [79], and large document understanding [27, 28, 52]. With the development of Vision Large Language Models (VLMs), automatic processing of HRIs has been a promising direction [64]. As shown in Figure 1c, Gemini [59], Claude [4], and GPT [3] can support images exceeding 1k resolution, enabling a wide range of real-world applications, such as 24/7 street monitoring [79], galaxy research [35], and radiology analysis [40].

However, even though existing VLMs can allegedly handle inputs of high-resolution images, there is a lack of comprehensive HRI benchmark, hindering objective calibration and measurement of progress on the effectiveness of HRI understanding. First, HRI evaluation is often missing from the official reports of mainstream VLMs. Figure 1c lists most of the vision-based benchmarks that VLMs are evaluated on, such as MMMU [76], VQAv2 [23], and AI2D [37]. Their average resolution is typically below 1k, making them unsuitable for HRI evaluation. Moreover, there is no comprehensive real-world or diagnostic benchmark for HRIs. As shown in Table 1, the existing real-world datasets with HRIs tend to either focus on specific scenarios, like long-range images [65], or particular resolutions, such as 8k [79]. The current diagnostic evaluation, namely, Multi-modal Neelde-in-the-Haystack (NIAH), primarily focuses on long text [48] or a mixture of low-resolution images [67].

To address this gap, we introduce HRScene, a unified benchmark for HRI understanding, covering diverse real-world scenes. HRScene incorporates 25 real-world tasks with resolutions ranging from 1024 × 1024 to 35,503 × 26,627, and 2 synthetic diagnostic datasets with 1k to 4k resolution. As shown in Figure 1a, we propose a task taxonomy to guide the development of HRScene : (1) we identify 8 categories of HRI tasks: Daily pictures, Urban planning, Paper scanned images, Artworks, Multi-subimages, Remote sensing, Medical Diagnosing, and Research understanding. (2) We focus on the 25 real-world scenes distributed across



(a) (b) (c)

Figure 1. (a) Overview taxonomy of the HRScene. (b) Performance of some VLMs on HRScene. (c) Comparison between the benchmarks that the mainstream VLMs are evaluated on and HRScene. The y-axis is the $\sqrt{\text{total pixel}}$. The boxes/icons indicate the image resolution they contain/support. The black lines inside each box show the average resolutions.

these categories, such as street monitoring and medical image understanding. (3) Each scene evaluates diverse capabilities of LLMs, such as counting, temporal and semantic reasoning, holistic image judgment, visual retrieval, spatial relations, and small object detection. HRScene is collected from 25 existing data resources, and 8 of them are re-annotated by 10 graduate-level annotators, with diverse view scales, ranging from microscope to radiology, street views, long-range, and telescope images. It contains high-resolution images of real objects, electronic documents, and composite multi-subimages. Besides, six datasets require domain-expert knowledge, while the remaining 19 belong to general domains. The diagnostic dataset is synthesized by combining the target image with the gold answer and visually similar distractors arranged in different orders to assess HRI utilization. Overall, HRScene comprises 7,068 images, with 2,008 of them being re-annotated.

Table 1. Comparison with existing real-world benchmarks and Multi-modal NIAH diagnosis.

Benchmark	Scenes	Easy Eval	Highest Res	Avg Res
MME [79]	5	✓	5304 × 7952	2000 × 1500
HR-Bench [65]	1	✗	8000 × 8000	8000 × 8000
HRScene (ours)	25	✓	35503 × 26627	4828 × 4078
NIAH	High-Res	Multi-Res	Real Img	Needle in
MM-NIAH [67]	✗	✗	✓	text
MileBench [58]	✗	✗	✓	text
Visual Haystack [71]	✗	✗	✗	image
HRScene (ours)	✓	✓	✓	image

We conduct extensive experiments to evaluate the HRI understanding of 28 popular VLMs, including 6 proprietary VLMs: GPT [32], Claude [4], and Gemini [60] families, and 22 open-sourced models, including InternVL2 [12], DeepSeek [72], Phi [2], Qwen [63], MolMo [15], LLava [47], and Llama [19] families. As shown in Figure 1b, experiments on real-world tasks demonstrate that current VLMs perform modestly, with an average accuracy of around 50%, highlighting substantial challenges of HRScene. Besides, we also provide the human performance of all real-world datasets by engaging graduate-level annotators to annotate 750 image-question pairs. Our synthetic datasets provide a fine-grained understanding of VLM performance, revealing two robust issues across VLMs, including regional divergence and lost-in-the-middle, shedding light on future improvement directions.

Our contributions are: (1) we propose HRScene, a unified benchmark with 25 real-world and 2 diagnostic datasets; (2) we benchmark 28 models on HRScene and show the significant performance gap; (3) we discover two salient issues of VLMs, including regional divergence and the lost-in-the-middle.

2. Related Work

VLMs for High-resolution Image. Recent advances in vision-language models have demonstrated remarkable capabilities in understanding and reasoning about visual content [19, 32, 47, 59]. However, processing high-resolution images remains a significant challenge due to computational constraints and the need to capture both fine-grained details and global context [8]. Two main categories of approaches have emerged to address these challenges. The first category employs a dual encoder architecture that processes the high-resolution and low-resolution of the same image in parallel [25, 50, 51]. A low-resolution encoder, typically CLIP [55], captures coarse-grained features for global understanding, while a high-resolution encoder based on convolutional neural networks [49] or ob-

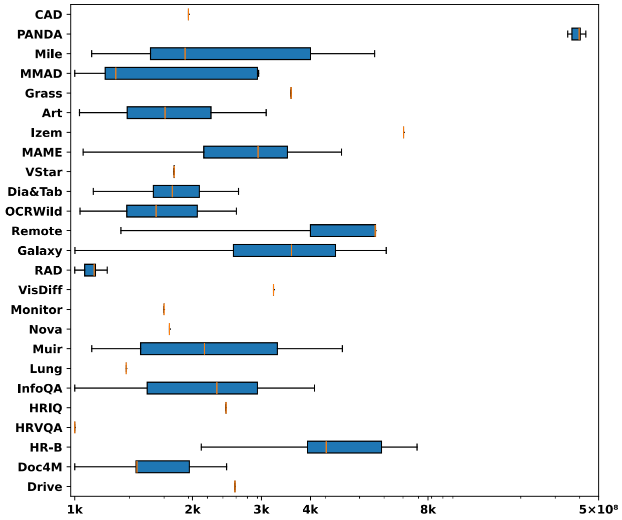


Figure 2. Distribution of resolution of each dataset. X-axis is the resolution and nk indicates the resolution is at least $n^2 * 10^6$ pixels.

ject segmentation models [38] preserves fine-grained details. The resulting high-resolution and low-resolution tokens are then fused or concatenated before being passed to large language models. This design achieves computational efficiency while leveraging both global and local visual features. The second category utilizes a splitting strategy [18, 24, 26, 30, 44, 53, 72]. A high-resolution image is first downsampled and processed by a vision encoder to capture the global context, while the original image is then divided into multiple sub-images processed by a vision encoder with the same resolution to extract local features. It enables efficient processing of high-resolution images while preserving fine details, proving particularly effective for tasks requiring detailed local analysis and global scene understanding. In this work, we propose HRScene, a benchmark with diverse scenes that challenge both global and local understanding capabilities of VLMs.

Evaluating High-resolution Image Understanding. High-resolution image understanding has been an important topic in computer vision [7]. Tradition research has usually focused on downstream tasks such as classification and detection, such as Crowd Counting [33, 34, 57, 68, 77], Autonomous Driving [14, 31, 43, 56, 75], Aerial Image Classification [13], and Pathology [1, 5, 61, 74]. With the development of VLMs, more benchmarks have been introduced based on the stronger capability of VLMs, these tasks usually incorporate text instructions, logical reasoning, and complex question answering. For instance, MME-Realworld [79] proposes five tasks with small objects for VLMs to recognize and answer the given question; MileBench focuses on the ability of VLMs to discern relationships of multiple images that are possibly temporal or semantically interconnected; and MuirBench contains 12 diverse multi-image tasks (e.g., scene understanding, ordering) with 10 multi-image relations. However, these works are separately proposed and focus on limited scenes, failing to evaluate the VLMs' capability to HRI understanding comprehensively.

Needle-in-the-Haystack. Needle in the Haystack (NIAH) test is initially proposed for long text input [48]. It is used to evaluate the LLM capability in long context understanding and mark the text regions that LLMs fail to understand. Recently, this NIAH test has been extended for Multi-modality settings. MM-NIAH [67] evaluates VLMs with a question on the synthetic sub-image embedded in a larger whole image. MileBench [58] mixes the text and image input and asks the model to retrieve both text and image information. Visual Haystack [71] feeds VLM with multiple images and tests the model's robustness to the permutation of this input. Different from these works, we analyze the HRI and permute the sub-images inside one large image.

Table 2. General Statistics of HRScene.

Samples	#	Tasks	#	Elements	#
Total	7068	Total	27	Images	7068
Reannotated	2005	Real-world	25	Questions	5807
Scratch	384	Synthetic	2	Options	34372

3. The HRScene Benchmark

This section outlines our benchmark construction process, including a brief overview of each adopted category and the manual efforts to ensure their adoption and maintain high data quality.

3.1. Collection Guidelines

As mentioned previously, there is a significant gap in the lack of unified, comprehensive, and easy-to-use HRI benchmarks for VLMs. HRScene is motivated to address this gap, offering a high-quality evaluation benchmark for HRI understanding, and pushing VLMs to a general-purpose HRI processor. To create a high-quality benchmark, we consider the following guidelines for the creation: (1) consider possible real-world scenes where users need VLMs to process HRIs. Think in broader categories and taxonomy. (2) Create comprehensive and most important tasks for each category without duplication of tested capability. (3) Ensure HRScene is easy to use. Each task does not need too many data points while being easy to verify the correctness.

The taxonomy for this work is introduced as follows: First, we identify 8 categories of scenes: Daily pictures, Urban planning, Paper scanned images, Artwork, Multi-sub-images, Remote sensing, Medical Diagnosing, and Research understanding. Second, we cover a wide array of

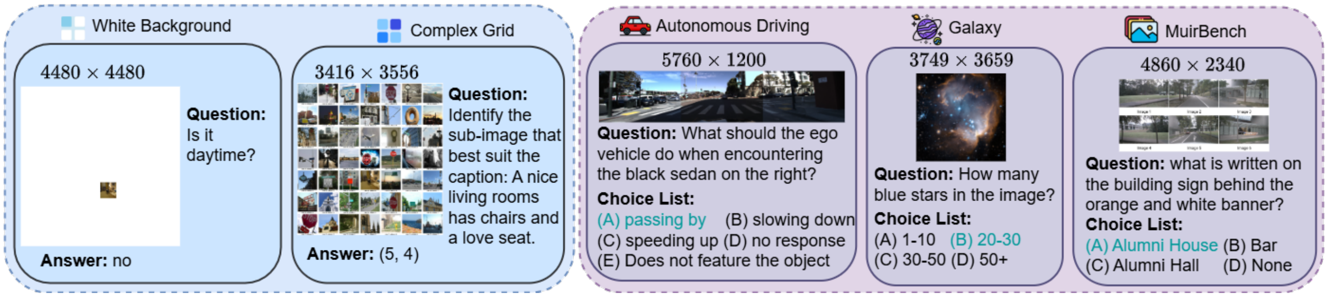

White Background
4480 × 4480
Question: Is it daytime?
Answer: no
Complex Grid
3416 × 3556
Question: Identify the sub-image that best suit the caption: A nice living rooms has chairs and a love seat.
Answer: (5, 4)
Autonomous Driving
5760 × 1200
Question: What should the ego vehicle do when encountering the black sedan on the right?
Choice List:
(A) passing by (B) slowing down
(C) speeding up (D) no response
(E) Does not feature the object
Galaxy
3749 × 3659
Question: How many blue stars in the image?
Choice List:
(A) 1-10 (B) 20-30
(C) 30-50 (D) 50+
MuirBench
4860 × 2340
Question: what is written on the building sign behind the orange and white banner?
Choice List:
(A) Alumni House (B) Bar
(C) Alumni Hall (D) None

Figure 3. Some examples of HRScene. Blue ones are diagnostic datasets and purple ones are real-world datasets.

camera types, from microscopes to daily cameras, long-range cameras, x-ray devices, remote sensing cameras, and telescopes. Last but not least, we enrich the format and requirements of the datasets by including both single-image and multi-image data, expert and non-expert level QA, as well as small object detection and global image understanding.

3.2. Data Collection

In general, for all tasks, we ensure the selection of a balanced subset by uniformly sampling across all categories, such as question types, subtasks, and domains. We also filter out any images that do not meet the minimum resolution threshold of 1024x1024 (1K). After sampling, all datasets undergo human inspection, and any low-quality data points are removed. The supplementary materials contain more details about all datasets. **Daily pictures** include the pictures captured by daily users with high-resolution cameras, including *HRIQ* [29], designed for Blind Image Quality Assessment, *VStar* [70], designed for image search of objects in daily images, *MMAD* [36], a dataset for abnormal detection of daily items, and *HR-Bench* [66], for long-range picture question answering. **Urban planning** contains the scenes that are helpful for intelligent urban construction: *Autonomous Driving* [79] contains images from car camera, *Monitoring* [79] contains street camera images, *OCR-Wild* [79] contains OCR of street signs, and *PANDA* [68] is the monitor images from a crowded environment. **Paper scanned images** consist of scanned images from paper or documents, including structural documents (*DocStruct4M* [27]), rich graphic documents (*InfoVQA* [52]), chart data (*NovaChart* [28]), and complex diagrams and tables (*Diagram&Table* [79]). **Artwork** tests the model's capability to understand art or design. *MAME* [54] contains the artwork images from museums, *ArtBench* [45] is a dataset with various paintings of different themes. *CAD* [21] is a floor plan dataset for interior design. **Multi-sub-images** contain images composed of multiple small sub-images. This includes *VisDiffBench* [20] that requires telling the difference between two groups of images. *MileBench* [58] contains frames from a video to form a larger image. *MuirBench* [62] focuses on the multi-image of various scenes and tests diverse capabilities, such as spatial relations. **Remote sensing** is geographic images captured by remote devices. This category includes *RemoteSen* [79] for aerial image QA and *HRVQA* [42] for remote small object detection. *Animal* [69] contains aerial images of geese from Izembek Lagoon in Alaska. **Medical Diagnosing** is for medical purposes. *VQA-RAD* [39] is a VQA dataset for radiology images of various types (X-rays and CT scans). *LungHist700* [17] is a collection of histopathological lung tissue images. **Research** includes the images for expert-domain research. *Grass* [6] is phenological stage classification and raceme counting of Urochloa images. *Galaxy*[1] contains the images from Hubble telescopes.

3.3. Data Reannotation

Task annotation Although most of the datasets contain annotations that can be directly used, some of them are not easy to evaluate or do not have labels. To this end, we ask 10 graduate-level annotators to reannotate 8 datasets. For six of them, we construct several wrong options so that each sample has at least 4 options in total. The wrong options are designed to be distracting and valid. Also, for two of the datasets, we annotate from scratch, with one question-answer pair for each image. For datasets with numeric answers, we automatically generate random numbers $r \in [-a, +a]$ multiple times to ensure 4 options, where a is the correct answer.

Human performance annotation. After all task annotations are done, we further collect their human performance. We pick 30 samples from each dataset and assign these samples to annotators to generate answers. We use this answer to compute the human performance. We also ask for feedback and comments from the annotator. If the annotator raises any concerns about the dataset, we revise its construction until the samples resolve the annotator's concerns.

[1] https://esahubble.org/images/

Table 3. Overall performance of all models on real-world datasets of HRScene. The models are grouped according to the parameter sizes. **Bold** indicates global best performance, while underline represents the best of the group. Avg. is the mean value of the column/row. Each category represents the average score of every sample of the corresponding category in the *test* set.

Model	Avg. *testmini*	Art	Daily	Medical	Paper	Remote	Research	Sub-Img	Urban	Avg.
Random	–	15.32	26.65	22.33	23.44	22.38	21.12	25.42	21.25	22.46
Phi3.5 4B	44.78	57.32	52.36	32.66	55.39	37.87	40.84	57.75	35.74	47.35
DeepSeek-Janus 7B	38.65	53.78	41.44	35.88	36.00	35.82	37.36	47.17	34.48	40.19
MolMo 7B-D	44.43	47.89	54.68	31.16	52.47	42.73	31.70	52.97	38.16	45.95
InternVL2 1B	35.58	31.07	36.83	30.17	32.37	30.34	26.22	37.62	25.59	31.63
InternVL2 2B	39.43	49.60	45.34	32.58	45.72	39.60	27.58	42.53	36.27	41.47
InternVL2 4B	47.46	60.74	47.18	42.02	57.91	35.96	35.52	66.16	40.77	49.23
InternVL2 8B	49.61	61.32	53.84	40.45	58.20	37.48	40.67	64.87	38.16	50.29
Qwen2-VL 7B	55.39	69.46	64.20	40.40	64.62	50.60	36.69	71.42	40.17	56.65
Llava-HR 7B	30.86	39.58	41.18	27.36	31.47	35.56	27.06	36.62	30.64	34.56
Llava-Next 8B	40.15	47.44	47.54	32.84	39.84	46.91	28.41	54.74	33.31	42.34
Llava-Next 13B	40.06	43.11	46.02	36.53	46.28	34.31	30.40	54.35	33.74	41.47
Llama3.2 11B	49.64	65.62	49.51	41.85	59.45	43.99	42.65	59.16	38.19	50.71
InternVL2 26B	53.86	66.97	57.95	38.15	64.62	42.58	36.04	68.27	**45.90**	54.69
DeepSeek-VL2 27B	53.09	71.83	58.78	35.86	61.84	46.16	34.88	66.05	44.59	54.73
InternVL2 40B	60.55	74.35	62.67	38.10	70.89	44.16	43.15	74.10	44.40	58.45
Llava-Next 34B	50.42	64.74	55.59	36.90	57.28	42.46	51.45	62.60	35.54	51.13
InternVL2 76B	55.38	69.74	61.68	38.08	67.07	43.50	29.40	73.68	41.28	55.64
Llama3.2 90B	54.88	70.66	51.90	40.02	64.31	44.85	31.56	63.48	39.28	52.60
Llava-Next 72B	47.00	61.00	54.20	36.15	57.19	42.03	31.56	59.13	34.74	48.69
Llava-OneVision 72B	55.96	68.26	64.64	42.00	68.91	46.18	52.12	68.75	40.28	57.45
MolMo 72B	54.02	60.30	58.16	35.91	63.67	52.16	**55.59**	64.97	43.99	55.12
Qwen2-VL 72B	**62.03**	75.85	**66.20**	43.69	78.13	**52.48**	39.36	**74.89**	44.66	**61.85**
Calude3 Haiku	40.84	57.90	37.14	27.66	55.08	30.05	29.24	57.43	27.67	41.34
Calude3.5 Sonnect	55.49	75.26	50.06	40.41	**78.85**	40.63	26.57	69.70	34.29	54.37
Gemini1.5 Pro	52.00	73.28	58.07	46.22	62.83	42.47	43.49	61.67	38.23	54.21
Gemini2.0 Flash	57.41	**76.46**	62.27	**51.94**	75.12	47.59	34.85	68.62	44.54	59.82
GPT-4o	51.24	69.13	55.90	22.63	66.80	44.05	35.38	65.13	41.72	52.91
GPT-4o mini	43.92	60.41	53.81	28.17	56.37	36.36	30.40	52.86	33.25	46.12
Avg.	48.72	61.54	53.18	36.64	58.17	41.75	36.08	60.60	37.84	49.68
Human	–	75.33	77.75	23.81	88.75	58.33	48.50	90.00	55.25	64.72

3.4. Data Synthesis

While real-world datasets provide a comprehensive evaluation of diverse scenes, a good diagnostic evaluation can point out the issues of VLMs, and further enhance the understanding of the model defects, guiding future research directions. To this end, we propose two diagnostic datasets with HRIs aiming to find the defect of VLMs on HRI understanding in two aspects.

WhiteBackground NIAH. This test aims to detect the regional defect of the VLMs, namely, Regional Divergence. Inspired by the NIAH test of long text [48] and the eye exam of humans, we propose to use a needle image to combine with a white background haystack to form an NxN grid (Figure 3). Specifically, we first prepare the image-question pairs from the VQAv2 dataset [23] and use the image as the *needle*. We then place it on different rows and columns of white girds as *haystack* and evaluate the differences in model performance.

ComplexGrid NIAH. This test is to diagnose the VLM capability of retrieving the correct image among multiple distracting images. Also inspired by the NIAH test in long text [48], we use an image search tool to extract the most similar images from the dev set of VQAv2. Then, we composite them to form a larger grid. We collect the caption of the needle and ask the model to point out the rows and columns where the needle is located (Figure 3).

3.5. Data Analysis

Table 2 and Figure 2 shows the statistics of the datasets. As shown in Figure 2, all of the datasets have a resolution higher than 1k. For most of the datasets, the resolution is between 1k and 8k while PANDA contains images with around 5×10^8 pixels, showing the diverse high-resolution distribution of HRScene. Figure 3 displays some examples of HRScene, the questions of HRScene cover a wide range, such as indexing the correct image, action prediction, counting, and text recognition. More examples can be found in Supplementary materials.

3.6. Data Preparation and Release

HRScene consists of 7068 samples, divided into three splits: *val*, *testmini*, and *test*. *val* contains 750 samples. These samples are identical to human-annotated ones, designed for fine-grained validation of the users' VLM settings. *testmini* comprises 1000 samples, picked from each HRScene real-world datasets, intended for rapid model development evaluation or for those with limited computing resources. To ensure *testmini* maintains a distribution closely resembling the whole set, we adopt this sampling: (1) first, randomly sample questions with a threshold number of 25 for each dataset; (2) then, randomly sample the remaining questions for each dataset according to its proportion in the entire set. The *test* features the remaining 5323 samples for standard evaluation. Notably, the answer labels for *test* will not be publicly released to facilitate fair evaluation. Instead, we maintain an online evaluation platform for user submissions. Evaluation results will be shown on an official leaderboard.

4. Experiments

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** [2], two **DeepSeek** [9], seven **InterVL2** [10–12, 22], two **Qwen2-VL** [63], two **MolMo** [15], one **LLaVA-Onevision** [46],one **LLaVA-HR** [51], four **Llava-Next** [78], two **Llama-3.2** [19], and two **GPT-4o** [32], two **Gemini** [60], and two **Claude** [4].

For WhiteBackground, we follow the accuracy in VQAv2[23]. 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. Supplementary materials show more details.

4.2. Overall Results on Real-world Datasets

Table 3 shows the comparison of all models on HRScene real-world datasets. To make the comparison fair, we cluster the open-sourced models into 3 groups with similar parameter sizes (Small: 1B-13B, Medium: 14B-34B, Large: >35B), and compare the models inside each group. As can be seen, the best small model is Qwen2-VL 7B, with an average performance of 56.65% on all 25 tasks. However, InternVL2 4B obtains group SOTA on Urban and Medical categories. For Medium, InternVL2 40B obtains the highest average score of 58.45%. While for Large models, Qwen2-VL 72B obtains the best performance again, with 61.85% on average. For proprietary models, the best model is Gemini2.0 Flash with 59.82% performance. Supplementary materials contain the full results of all models.

Comparing these four types of models (Open-sourced Small, Medium, Large, and Proprietary), we can observe that the performance increases with the increasing model sizes. However, the best model globally is Qwen2-VL 72B, the only model whose performance is above 60%, surpassing the GPT 4o, and Gemini. We further explore the reason by checking the dataset details. As shown in Table 5, 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. This result highlights **the importance of the HRI processing capability of native resolution to obtain high performance.**

On the other hand, for VLMs, the difficulties of each category vary. As shown in the second-to-last row of Table 3, Medical and Research obtain the lowest performance as it requires domain knowledge, such as pathology and medical images. Then, Remote and Urban planning is also challenging because it involves intensive counting and small object detection tasks. The simplest ones are Art and Sub-Img, which mainly focus on global understanding and reasoning capabilities. However, **the average performance across all categories is only 49.68%, showing the large gap between VLMs and efficient HRI processing.**

4.3. Overall Results on Diagnostic Datasets

WhiteBackground. Table 4 shows the statistics of the WhiteBackground diagnosis. We report the average performance of the samples (Perf ↑), the performance drop with image size increasing from 1x1 (Size ↑), and the region expectation gap (Region ↓), which is the difference between the highest performance region and the mean performance of every region. We call this *Regional Divergence*. We propose to use these metrics because the model can be improved by (1) being more robust on image size extension, especially with simple white background fillings, and (2) inside each HRI, maintaining the same performance with each region, specifically being the same performance as the highest region. As shown in Table 4, most of the models cannot maintain consistent performance with increasing image size. Furthermore, models exhibit significant Region Divergence, usually amplified with increasing image size. For instance, Gemini-2.0-Flash obtains 39.85% Divergence on 10x10 grids, meaning that if the answer to the question is located at a random region, the performance will be around 40% lower than the best one among 100 regions.

Table 4. Diagnostic NIAH test on WhiteBackground dataset, **bold** indicates the best performance.

	1x1	3x3			5x5			7x7			10x10		
	Perf ↑	Perf ↑	Size ↑	Region ↓	Perf ↑	Size ↑	Region ↓	Perf ↑	Size ↑	Region ↓	Perf ↑	Size ↑	Region ↓
InterVL2 - 1B	62.46	60.17	-2.29	25.16	55.72	-6.74	34.12	52.29	-10.17	39.37	53.11	-9.35	23.82
InterVL2 - 2B	65.66	63.91	-1.75	24.44	60.29	-5.37	31.63	56.77	-8.89	35.22	49.76	-15.90	40.03
InterVL2 - 4B	62.06	62.45	**0.39**	27.01	58.19	-3.87	35.34	54.60	-7.46	39.15	56.54	-5.52	19.45
InterVL2 - 8B	68.66	66.10	-2.56	24.69	62.66	-6.00	30.87	58.86	-9.80	35.93	53.44	-15.22	22.42
Phi - 4B	77.46	74.00	-3.46	10.06	69.64	-7.82	17.02	65.71	-11.75	20.54	61.44	-16.02	21.28
Qwen2-VL - 7B	85.93	84.22	-1.71	5.30	83.14	-2.79	**6.52**	81.71	-4.22	7.55	79.91	-6.02	10.56
LLaMa3.2 - 11B	72.53	69.14	-3.39	15.72	66.11	-6.42	19.75	61.73	-10.80	27.13	41.17	-31.36	**3.84**
InterVL2 - 26B	82.60	80.87	-1.73	6.45	77.43	-5.17	11.03	74.63	-7.97	13.29	65.41	-17.19	16.99
InterVL2 - 40B	84.53	83.42	-1.11	**4.57**	80.02	-4.51	8.84	78.16	-6.37	11.29	74.95	-9.58	13.18
InternVL2-Llama3-76B	85.33	83.74	-1.59	4.59	77.09	-8.24	9.63	–	–	–	75.64	-9.69	13.35
DeepSeek-VL2-27B	72.06	49.71	-22.35	15.75	34.29	-37.77	23.37	28.41	-43.65	25.72	23.95	-48.11	23.30
LLava-Onevision-72B	**87.73**	**84.51**	-3.22	5.14	**84.04**	-3.69	10.40	–	–	–	–	–	–
Qwen2-VL - 72B	84.13	**84.51**	0.38	5.62	**84.04**	**-0.09**	6.62	**84.20**	**0.07**	**6.65**	**84.56**	**0.43**	9.61
LLaMa3.2 - 90B	75.46	72.69	-2.77	10.88	–	–	–	–	–	–	–	–	–
GPT-4o-mini	68.66	60.69	-7.97	13.77	52.53	-16.13	19.59	44.35	-24.31	25.64	32.94	-35.72	33.65
Claude-3-Haiku	62.60	59.92	-2.68	17.00	54.05	-8.55	26.61	49.89	-12.71	31.84	44.83	-17.77	36.89
Gemini-2.0-Flash	69.86	67.45	-2.41	6.75	67.35	-2.51	16.00	66.50	-3.36	25.64	63.86	-6.00	39.85

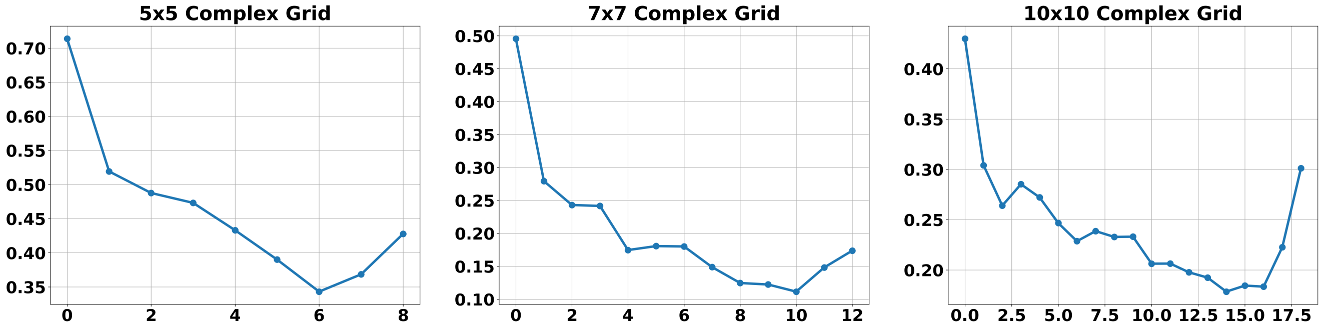


Figure 4. Performance of the regions averaged across all dataset points and all 18 VLMs. X-Axis is the Manhattan distance to the left upper corner, $|x-1|+|y-1|$ where x, y is the row and column of the needle image, while the y-axis is the performance of that sample. With the increase of the x-axis, the performance of the model exhibits a U-shape, with much lower performance in the middle. With the increase in the image size, the shape becomes more significant.

Table 5. Fine-grained comparison between Qwen 72B and GPT 4o on two datasets.

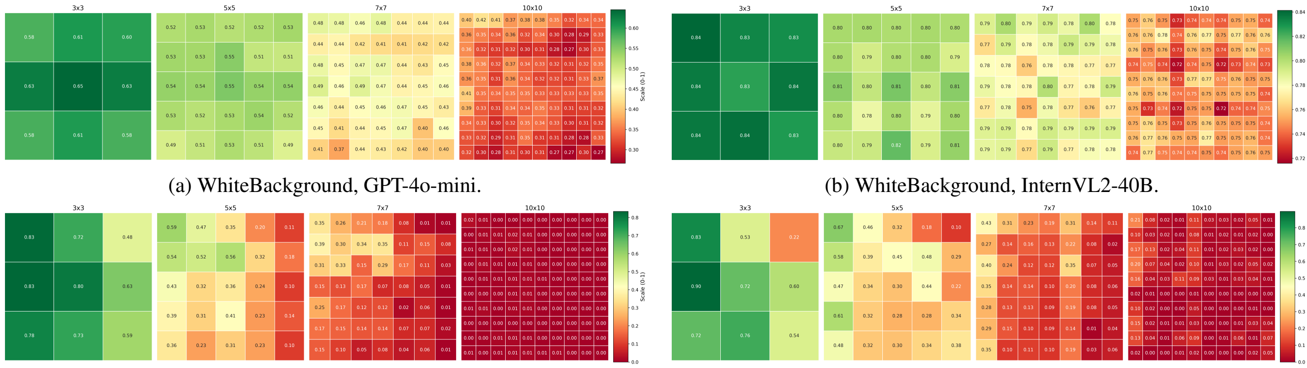
	HRVQA (1k)	Galaxy (20k)
Qwen 72B	69.59	**80.80**
GPT 4o	**73.82**	68.68

ComplexGrid. In the NIAH test [48], the authors found a significant drop when the gold answer is in the middle of a long context, namely lost-in-the-middle. This is also observed in multi-modal settings when the image is mixed with text [58]. Surprisingly, we discover a similar but non-identical behavior in HRI. Figure 4 shows the performance change of the models with increasing Manhattan distance from row 1, column 1 to the needle image. For instance, if the needle image is row 2, column 3, the Manhattan distance is computed as (2-1)+(3-1)=3. We observe a phenomenon that is similar to lost-in-the-middle. Differently, we observe the performance forms a U-shape based on the Manhattan distance from the left upper corner rather than the linear depth of the needle in traditional NIAH. We demonstrate that lost-in-the-middle-manhattan is novel and from the original lost-in-the-middle in supplementary materials.

5. Analysis

5.1. Influence of Model Size

We analyze the influence of model parameters on the performance. We plot the relation between VLMs' parameter size and average performance on 25 datasets. Next, we draw the trend line to fit the performance change. As shown in Figure 6, although different families of models' performances are different, their trends are all log-like increasing. This shows that (1) increasing the model size can effectively in-



(a) WhiteBackground, GPT-4o-mini.

(b) WhiteBackground, InternVL2-40B.

(c) ComplexGrid, Claude-3.5-Haiku.

(d) ComplexGrid, Phi-35-Vision.

Figure 5. Detailed performance of some models on two diagnose datasets.

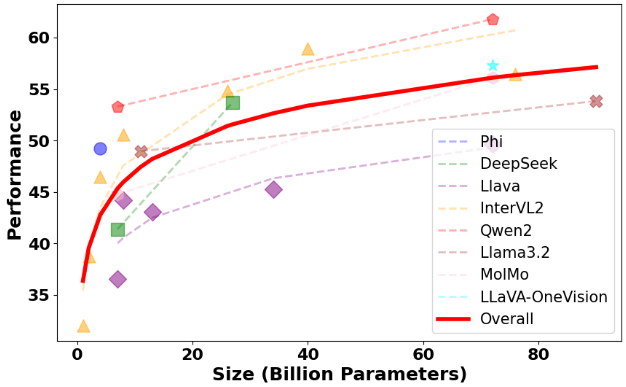


Figure 6. The relationship between model performance and model parameter size.

crease the HRI understanding, especially for the small models; and (2) the effect of increasing size is slowing down.

5.2. Global and Local Perception Trade-off

Processing HRIs requires capturing both fine-grained details and global context [8]. When an image becomes larger, it becomes more difficult to capture global information because of the larger amount of patches in the input. At the same time, it is too difficult to recognize objects when the image becomes too small. To evaluate this trade-off in HRScene, we run the Qwen2-VL 7B on two datasets: Autonomous Driving, and HR-Bench. We resize the image to 90%, 70%, 50%, 30%, and 10% to evaluate the same image with different sizes. Results in Figure 7 show that resizing to 1840 can obtain the best performance of Autonomous Driving, and 3628 for HR-Bench. This **indicates the trade-off between local and global information can lead to an optimal point that is different in different** datasets.

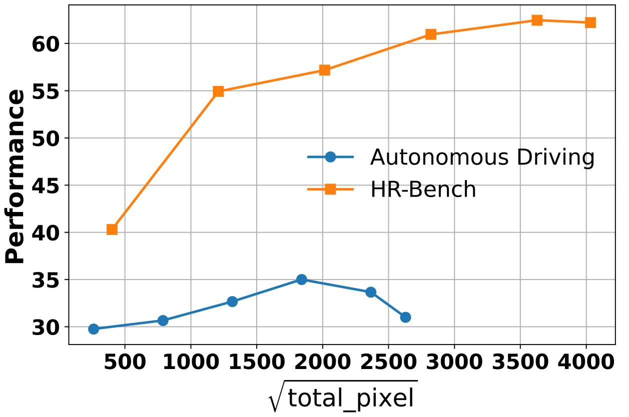


Figure 7. Performance change with image resizing on Qwen2-VL 7B.

5.3. Ablation on Multi-image Combination

For ComplexGrid dataset, multiple images are combined to form a large image with NxN grids. Thus, we conduct an ablation study on different combination methods to avoid unnecessary errors due to the non-optimal methods. We test 4 settings, dense is to combine all images without any interval between images or index text below each image. Then, we add an index to each image and a white interval between images. We evaluate four models on the 10x10 dataset 3. Results are in Table 6. Results show that indexing and interval are useful for the model to recognize the different sub-images and avoid errors in counting the index. Thus, we use both to construct the ComplexGrid dataset.

5.4. Fine-grained Analysis of Diagnostic Datasets

To further analyze the performance details of two diagnostic datasets, we draw a heatmap for the models. Figure 5 shows the performance divergence on different regions, where each grid represents the performance when

Table 6. Ablation study on different image combination methods.

	Qwen2-VL		InterVL2	
	7B	72B	8B	26B
dense	1.53	3.17	2.78	1.82
w/ index	8.22	18.33	17.15	6.30
w/ interval	2.33	5.17	4.28	1.99
w/ both	**8.62**	**18.72**	**20.54**	**8.28**

the needle image is on that grid. The results show that for the WhiteBackground dataset, the performance of different grids varies. Although different models do not have a unified pattern, the Regional Divergence is still significant, especially on larger images. For ComplexGrid, the results clearly show the lost-in-middle phenomenon with the increasing Manhattan distance, where the performance is the best at the upper left corners and gradually becomes worse with increasing Manhattan distance.

6. Conclusion and Future Work

In this paper, we propose HRScene, a unified benchmark for HRI understanding, consisting of 25 real-world datasets and two diagnostic datasets. Results show that the models exhibit low performance on real-world tasks, showing the challenge of HRScene, and display regional divergence and lost-in-the-middle on diagnostic datasets that can direct future improvement. In the future, researchers can develop high-performance, general-purpose HRI processors by fine-tuning on synthetic datasets or explore the deeper reasons for utilization issues. After developing a stronger model, it can be tested on real-world datasets of HRScene and submitted to the leaderboard of HRScene to obtain a direct comparison with other models on real-world scenarios.

References

[1] Brigham & Women's Hospital & Harvard Medical School Chin Lynda 9 11 Park Peter J. 12 Kucherlapati Raju 13, Genome data analysis: Baylor College of Medicine Creighton Chad J. 22 23 Donehower Lawrence A. 22 23 24 25, Institute for Systems Biology Reynolds Sheila 31 Kreisberg Richard B. 31 Bernard Brady 31 Bressler Ryan 31 Erkkila Timo 32 Lin Jake 31 Thorsson Vesteinn 31 Zhang Wei 33 Shmulevich Ilya 31, et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418): 61–70, 2012. 3

[2] 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, 6

[3] 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. 1, 3

[4] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024. 1, 2, 6

[5] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019. 3

[6] Darwin Alexis Arrechea-Castillo, Paula Espitia-Buitrago, Ronald David Arboleda, Luis Miguel Hernandez, Rosa N. Jauregui, and Juan Andrés Cardoso. High-resolution image dataset for the automatic classification of phenological stage and identification of racemes in urochloa spp. hybrids. *Data in Brief*, 57:110928, 2024. 4

[7] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. Efficient high-resolution deep learning: A survey. *ACM Computing Surveys*, 56(7):1–35, 2024. 1, 3

[8] Arian Bakhtiarnia, Qi Zhang, and Alexandros Iosifidis. Efficient high-resolution deep learning: A survey. *ACM Comput. Surv.*, 2024. 2, 8

[9] 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. 6, 2

[10] 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. 6, 2

[11] 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.

[12] 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, 6

[13] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 3

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[15] 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, 6

[16] Jorge Diosdado, Pere Gilabert, Santi Seguí, and Henar Borrego. LungHist700: A dataset of histological images for deep learning in pulmonary pathology. *Scientific Data*, 11 (1):1088, 2024. 1

[17] Jorge Diosdado, Pere Gilabert, Santi Seguí, and Henar Borrego. LungHist700: A dataset of histological images for deep learning in pulmonary pathology. *Scientific Data*, 11 (1):1088, 2024. 4

[18] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k HD. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 3

[19] 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. 2, 6

[20] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208, 2024. 4

[21] Zhiwen Fan, Lingjie Zhu, Honghua Li, Xiaohao Chen, Siyu Zhu, and Ping Tan. Floorplancad: A large-scale cad drawing dataset for panoptic symbol spotting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10128–10137, 2021. 4

[22] 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. 6, 2

[23] 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. 1, 5, 6

[24] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: An LMM perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[25] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for GUI agents. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024. 2

[26] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) Findings*, 2024. 3

[27] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024. 1, 4

[28] Linmei Hu, Duokang Wang, Yiming Pan, Jifan Yu, Yingxia Shao, Chong Feng, and Liqiang Nie. Novachart: A large-scale dataset towards chart understanding and generation of multimodal large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3917–3925, 2024. 1, 4

[29] Huang Huang, Qiang Wan, and Jari Korhonen. High resolution image quality database. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3105–3109. IEEE, 2024. 4

[30] Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Alleviating the semantic sawtooth effect for lightweight MLLMs via complementary image pyramid. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[31] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018. 3

[32] 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. 2, 6

[33] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013. 3

[34] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018. 3

[35] Raza Imam, Mohammed Talha Alam, Umaima Rahman, Mohsen Guizani, and Fakhri Karray. Cosmoclip: Generalizing large vision-language models for astronomical imaging. *arXiv preprint arXiv:2407.07315*, 2024. 1

[36] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: The first-ever comprehensive benchmark for multimodal large language models in industrial anomaly detection. *arXiv preprint arXiv:2410.09453*, 2024. 4

[37] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is

worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 1

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[39] J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. 4

[40] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 1

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2

[42] Kun Li, George Vosselman, and Michael Ying Yang. Hrvqa: A visual question answering benchmark for high-resolution aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 214:65–81, 2024. 4

[43] Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 473–488. Springer, 2020. 3

[44] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024. 3

[45] Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*, 2022. 4

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 6, 3

[47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2

[48] 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. 1, 3, 5, 7, 2

[49] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022. 2

[50] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2

[51] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu 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. 2, 6

[52] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 1, 4

[53] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. MM1: methods, analysis and insights from multimodal LLM pre-training. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[54] Ferran Parés, Anna Arias-Duart, Dario Garcia-Gasulla, Gema Campo-Francés, Nina Viladrich, Eduard Ayguadé, and Jesús Labarta. The mame dataset: on the relevance of high resolution and variable shape image properties. *Applied Intelligence*, 52(10):11703–11724, 2022. 4

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[56] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 3

[57] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2594–2609, 2020. 3

[58] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024. 2, 3, 4, 7

[59] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[60] 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. 2, 6, 3

[61] Mitko Veta, Yujing J Heng, Nikolas Stathonikos, Babak Ehteshami Bejnordi, Francisco Beca, Thomas Wollmann, Karl

Rohr, Manan A Shah, Dayong Wang, Mikael Rousson, et al. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *Medical image analysis*, 54: 111–121, 2019. 3

[62] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 4

[63] 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, 6

[64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[65] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *arXiv preprint*, 2024. 1, 2

[66] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *arXiv preprint arXiv:2408.15556*, 2024. 4

[67] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37: 20540–20565, 2025. 1, 2, 3

[68] Xueyang Wang, Xiya Zhang, Yinheng Zhu, Yuchen Guo, Xiaoyun Yuan, Liuyu Xiang, Zerun Wang, Guiguang Ding, David Brady, Qionghai Dai, et al. Panda: A gigapixel-level human-centric video dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3268–3278, 2020. 3, 4

[69] E.L. Weiser, P.L. Flint, D.K. Marks, B.S. Shults, H.M. Wilson, S.J. Thompson, and J.B. Fischer. Aerial photo imagery from fall waterfowl surveys, izembek lagoon, alaska, 2017-2019. U.S. Geological Survey data release, 2022. 4

[70] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 4

[71] Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *arXiv preprint arXiv:2407.13766*, 2024. 2, 3

[72] 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. 2, 3

[73] 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. 2

[74] Shuyi Yang, Longquan Jiang, Zhuoqun Cao, Liya Wang, Jiawang Cao, Rui Feng, Zhiyong Zhang, Xiangyang Xue, Yuxin Shi, and Fei Shan. Deep learning for detecting corona virus disease 2019 (covid-19) on high-resolution computed tomography: a pilot study. *Annals of translational medicine*, 8(7):450, 2020. 3

[75] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 3

[76] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 1

[77] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016. 3

[78] 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. 6, 3

[79] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. 1, 2, 3, 4

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

Supplementary Material

7. 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 sampels 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 perception 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. NxN 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 Divergence. 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 [41] 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.

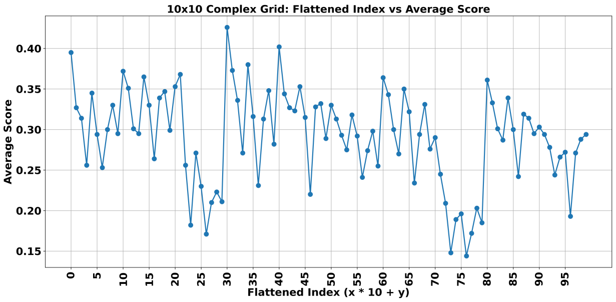


Figure 8. The performance of all models with an increase of patch id. Unlike lost-in-the-middle, no significant pattern can be observed.

8. Ablation on Lost-in-the-Middle

To test whether the observed U-shape is a trivial extension of existing work [48], 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 [63] 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 8. As shown, no significant patter 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.

9. Model Details

We include a total of 28 models in our experiment. **Phi-3.5** [2] is a lightweight model designed for efficient language understanding and generation. We include **Phi 3.5 vision instruct** [2] for experiments. **DeepSeek Janus Pro 7B** [9] is a model that integrates multi-modal reasoning capabilities. **DeepSeek-VL2** [73] is a vision-language model, with **deepseek vl2 27B** included in our evaluation. **InterVL2** [10–12, 22] 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** [63] is a vision-language model, and we consider both **Qwen2 VL 7B Instruct** and **Qwen2 VL 72B Instruct**. **MolMo** [15] is a series of models designed for molecular and scientific

Table 7. Overview of 25 real-world datasets and their statistics. * indicates that the dataset is reannotated.

Dataset Name	Explaination	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

applications. We include **Molmo 72B 0924** and its distilled variant, **Molmo 7B D 0924**. **LLaVA-Onevision** [46] is an open-source multimodal LLM, we selected **llava-onevision-qwen2-72b-ov-hf** model for our experiments. **Llava-Next** [78] 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** [3]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** [60]. **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**.

10. Performance Details on Real-world Datasets

Table 8 and Table 9 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 8. 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 Sonnect	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 9. 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 Sonnect	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

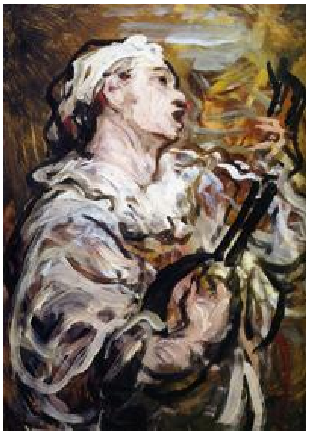
11. Prompts and Metrics

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.

12. Examples

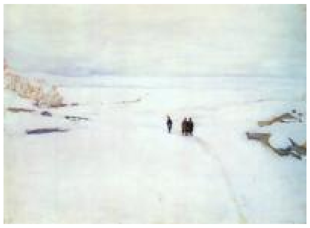
Table 10 to 32 show examples from HRScene real-world datasets. We compress the images to display them in the paper.

Table 10. 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

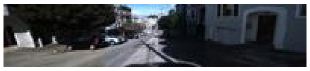
Answer: H



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

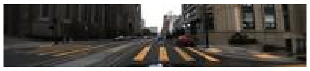
Answer: E

Table 11. 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

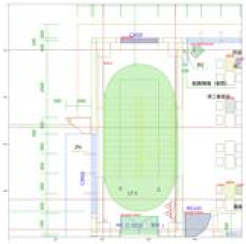
Answer: B



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

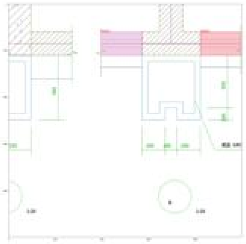
Answer: E

Table 12. Example from HRScene – CAD



How many doors are there in the image?
(A) 1
(B) 0
(C) 2
(D) 3

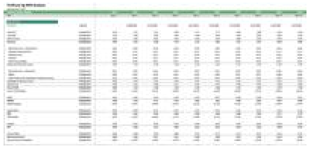
Answer: A



What is the shape of the shadow at upper left corner of the image?
(A) L-shape
(B) Oval
(C) Circle
(D) Squre

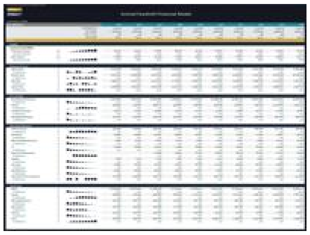
Answer: A

Table 13. 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 14. Example from HRScene – DocStruct4M

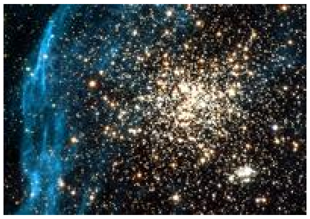
Read the following text: <doc>CALL FOR NOMINATIONS
BILINGUAL INSTRUCTIONAL ASSISTANT OF THE YEAR AWARD
[omitted]
Which of the following options is correct?
(A) the nominee's outstanding [omitted] 14 </doc>
(B) the nominee's outstanding [omitted] 14 </doc>
(C) the nominee's outstanding [omitted] 14 </doc>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

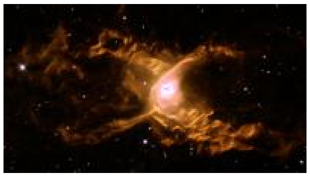
Answer: C

Table 15. 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

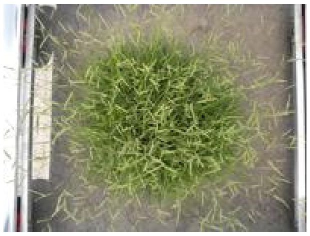
Answer: B



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

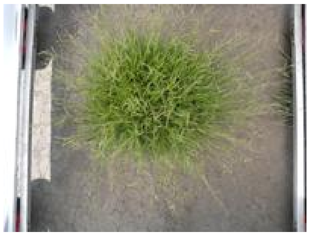
Answer: C

Table 16. 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

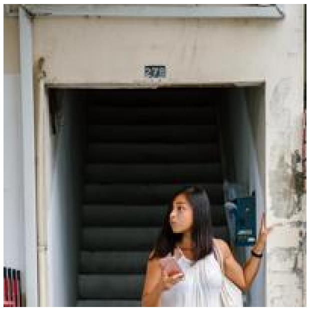
Answer: A



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)

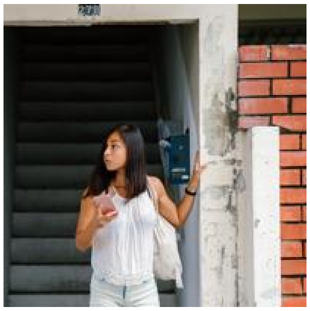
Answer: B

Table 17. Example from HRScene – HR-Bench



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

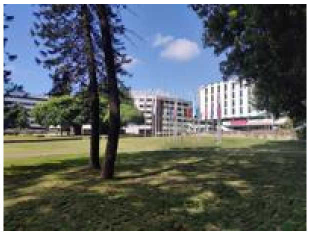
Answer: D



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

Answer: D

Table 18. 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 metadata or external references - your judgment should be based purely on visual quality.

(A) 1 bad

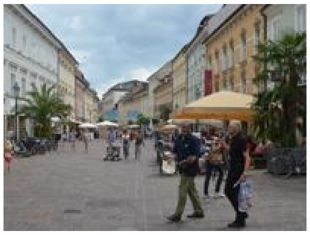
(B) 2 poor

(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 metadata or external references - your judgment should be based purely on visual quality.

(A) 1 bad

(B) 2 poor

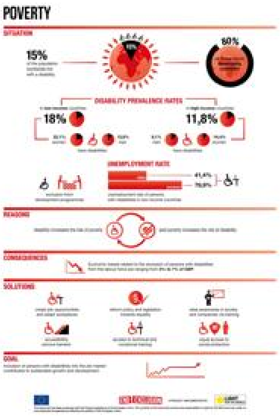
(C) 3 fair

(D) 4 good

(E) 5 excellent

Answer: D

Table 19. Example from HRScene – InfographicVQA



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

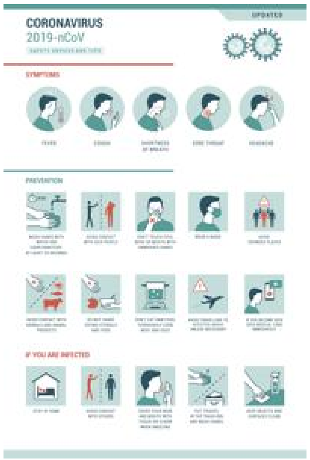
(A) '80, '80%'

(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'

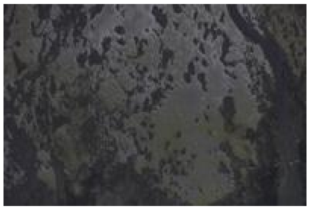
(B) 'cow, human'

(C) 'cat, cow'

(D) 'plane, apple'

Answer: A

Table 20. 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

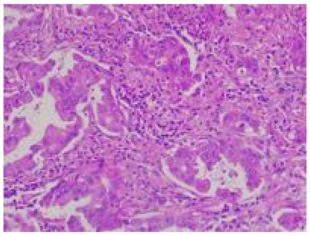
(B) 300-400

(C) 200-300

(D) 100-200

Answer: C

Table 21. 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)

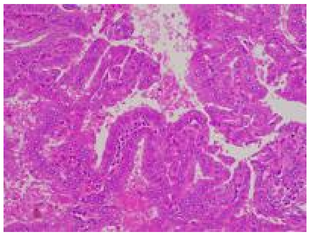
(D) Adenocarcinoma (Poorly differentiated)

(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)

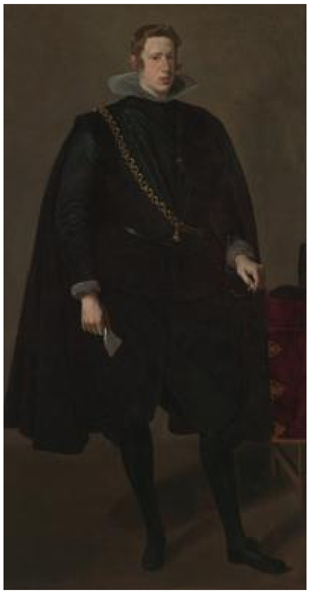
(E) Squamous cell carcinoma (Well-differentiated)

(F) Squamous cell carcinoma (Moderately differentiated)

(G) Squamous cell carcinoma (Poorly differentiated)

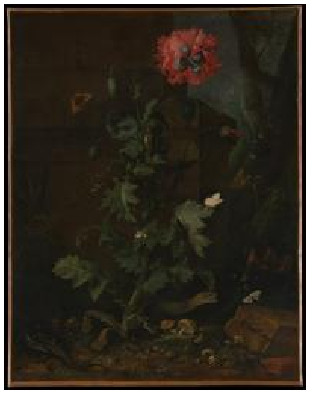
Answer: B

Table 22. 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

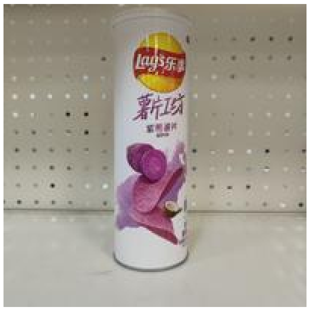
Answer: J



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

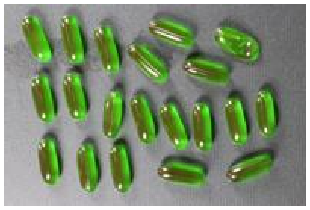
Answer: B

Table 23. 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

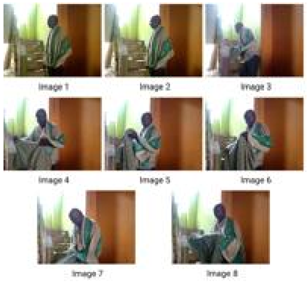
Answer: C



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.

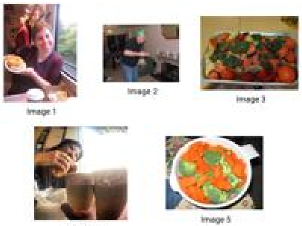
Answer: A

Table 24. 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

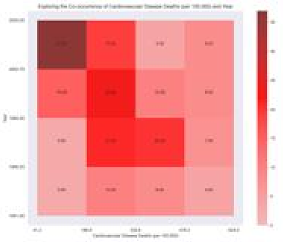
Answer: C



<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>

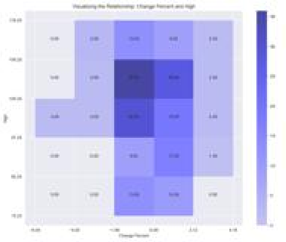
Answer: C

Table 25. 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

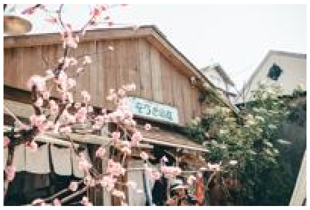
Answer: A



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

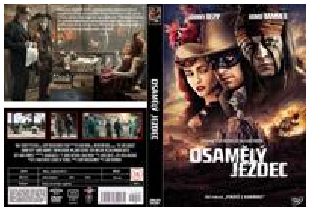
Answer: E

Table 26. Example from HRScene – OCR in the Wild



What is the content on the plaque in the center of the picture?
(A) 安らぎの梃
(B) 安らぎの廷
(C) 安らぎの挺
(D) 安らぎの庭
(E) This image doesn't feature the content.

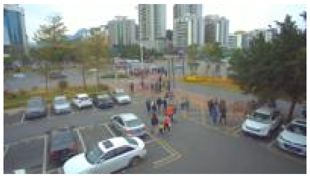
Answer: D



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.

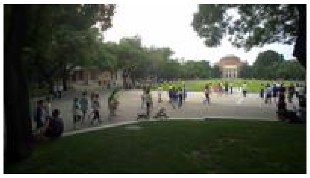
Answer: D

Table 27. Example from HRScene – PANDA



How many riding person(s) are in the image?
(A) 35
(B) 27
(C) 23
(D) 44

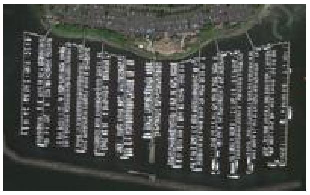
Answer: C



How many riding person(s) are in the image?
(A) 11
(B) 12
(C) 21
(D) 15

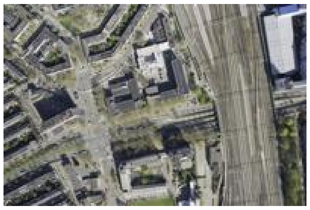
Answer: B

Table 28. 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.

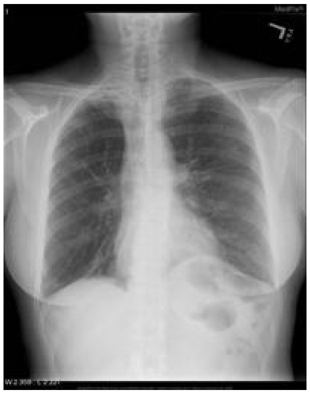
Answer: A



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.

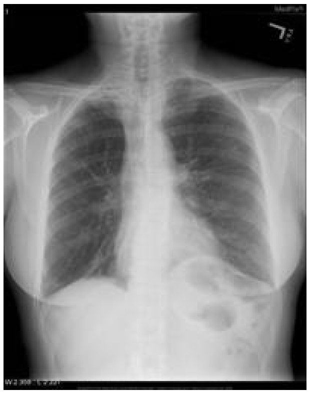
Answer: D

Table 29. Example from HRScene – VQA-RAD



Is the trachea midline?
(A) Yes
(B) No
(C) Not specified

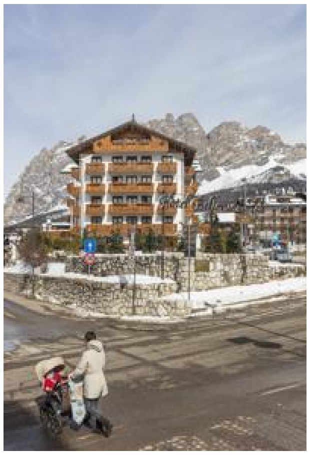
Answer: A



Is there evidence of an aortic aneurysm?
(A) Yes
(B) No
(C) Not specified

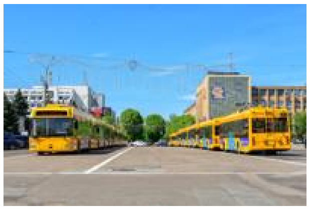
Answer: B

Table 30. 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.

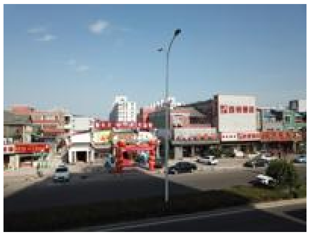
Answer: A



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.

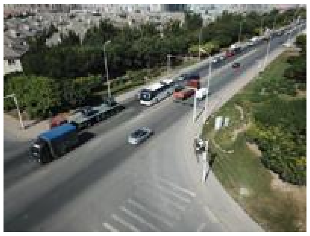
Answer: B

Table 31. 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

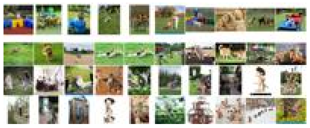
Answer: E



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

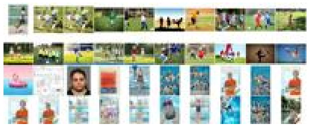
Answer: E

Table 32. 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