Vision-Language Models Can't See the Obvious

Yasser Dahou* Ngoc Dung Huynh* † Phuc H. Le-Khac
Wamiq Reyaz Para Ankit Singh Sanath Narayan

Technology Innovation Institute, Abu Dhabi, UAE
† Deakin University, Australia
https://salbench.github.io

Abstract

We present Saliency Benchmark (SalBench), a novel benchmark designed to assess the capability of Large Vision-Language Models (LVLM) in detecting visually salient features that are readily apparent to humans, such as a large circle amidst a grid of smaller ones. This benchmark focuses on low-level features including color, intensity, and orientation, which are fundamental to human visual processing. Our SalBench consists of images that highlight rare, unusual, or unexpected elements within scenes, and naturally draw human attention. It comprises three novel tasks for evaluating the perceptual capabilities of LVLM: Odd-One-Out Detection, Referring Odd-One-Out, and Visual Referring Odd-One-Out. We perform a comprehensive evaluation of state-of-the-art LVLM using SalBench and our findings reveal a surprising limitation: LVLM struggle to identify seemingly obvious visual anomalies, with even the advanced GPT-4o achieving only 47.6% accuracy on such a simple task. SalBench will be an important step in measuring the capabilities of LVLM that align with the subtle definition of human attention.

1. Introduction

Large Vision-Language Models (LVLM) have emerged as a central focus in recent computer vision research [4, 12, 31, 32, 44, 60]. The primary advantage of these models lies in their ability to reason about images using the Large Language Model (LLM) knowledge about the world, and solve tasks that surpass the capabilities of traditional vision models. LVLMs enable a better reasoning about the visual content using the prior world knowledge.

Current LVLM exhibit impressive performance on existing benchmarks [11, 17, 27, 34, 55, 57], which evaluate on a range of capabilities, from general visual question answering [17] to tasks requiring college-level subject knowledge and critical reasoning, such as MMMU [57]. With the main focus of these benchmarks being high-level complex tasks, it raises a fundamental question: can LVLM perform equally well on simple perceptual tasks, such as identifying a black dot on a white background? This aligns with Moravec's paradox [40], which suggests that high-level reasoning tasks are computationally simpler for artificial intelligence systems than low-level perceptual and sensorimotor skills. Consequently, we may find that LVLM excel at complex tasks present in existing benchmarks while strug-

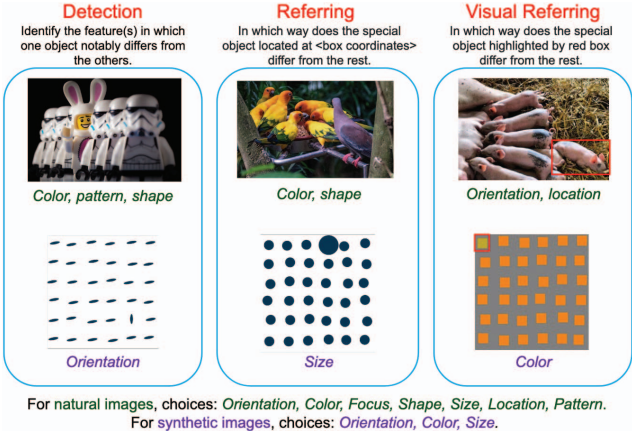

Figure 1. **Our SalBench evaluates the perceptual capabilities of vision-language models**. It comprises natural and synthetic images, which consist of a single salient target among multiple distractors. In natural images, the target can vary from the rest in orientation, color, focus, shape, size, location and pattern, while it differs only in orientation, color or size in synthetic images. SalBench evaluates the odd-one-out understanding under three tasks: Detection, Referring and Visual Referring. In detection task, the odd features must be directly predicted. Moreover, the box coordinates of the target are provided as context in text (referring task) or as a highlighted box in the image (visual referring task) for predicting the odd features of the target. Best viewed zoomed-in.

*Joint first authors
† This work was done when Ngoc Dung Huynh was intern at TII
Correspondence: yasser.djilali@tii.ae

gling with seemingly simple perceptual tasks that humans perform effortlessly. To this end, we propose a benchmark for quantifying the alignment of vision-language models on low-level perceptual tasks of the human attention.

While LVLM effectively capture high-level features such as cars and humans, they are likely to struggle to represent crucial aspects of human visual attention that have been extensively studied in neuroscience. Visual search, which is a fundamental process shaping human attention [24, 49] involves the brain's parallel processing of regions that differ significantly in one feature dimension, such as color, intensity, or orientation. These low-level features serve as basic mechanisms of the human visual system. By examining the LVLM performance on simple saliency-driven images, we aim to gain insights into the current state of these models relative to human perceptual abilities and identify areas for necessary improvements. Our key contributions are:

- We propose SalBench as an open-source benchmark, for evaluating and aiding the improvement of the perceptual capabilities of LVLM. To this end, we augment the P3/O3 datasets [24], which comprise of images with a single distinctive target among many similar distractors, with language instructions and create three novel tasks: Odd-One-Out Detection, Referring Odd-One-Out, and Visual Referring Odd-One-Out, as shown in Figure 1. Our benchmark aims to serve as a tool for assessing the progress in aligning LVLM with human visual attention.
- We conduct a comprehensive analysis of LVLM performance on SalBench, uncovering striking discrepancies between these advanced models and human visual capabilities. Our findings show that even state-of-the-art LVLM, including GPT-4o [42], struggle with basic saliency detection tasks that are trivial for humans.
- We provide insights into the limitations of LVLM in processing low-level visual features, highlighting the need for improved alignment between these models and human visual attention mechanisms. Our work demonstrates the importance of incorporating neuroscience principles into the development of future vision-language models.

2. Related Works

Many vision-language benchmarks have been introduced in the literature for evaluating and comparing LVLM on various tasks. The early benchmarks were mostly single-task oriented and can be broadly classified into captioning, general visual-question answering (VQA), and text-centric VQA. The captioning task is used to measure the LVLM's caption generation quality and is mostly evaluated on COCO [30] and NoCaps [3] benchmarks using BLEU, CIDER, ROUGE metric scores. Similarly, general VQA benchmarks like VQAv2 [17], GQA [21], ScienceQA [33], OK-VQA [44], VizWiz [18], Pope [28] typically ask general questions about the objects/scene in the image. Furthermore, text-centric VQA benchmarks such as OCRVQA [39], TextVQA [45], ST-VQA [8], DocVQA [36], ChartQA [35], InfoVQA [37], and AI2D [22] additionally focus on the vision-language model's ability to detect text in the image (Optical Character Recognition) and then answer the questions.

With the increasing capabilities of vision-language models in handling different types of tasks, the benchmarks have also evolved to evaluate the LVLM in a fine-grained manner. To this end, more recently, MMBench [32], MME [16], MM-Star [11], and MMMU [56] are curated to test the LVLM on a mix of different aspects, such as reasoning, perception, knowledge, chart interpretation, *etc*. Similarly, MathVista [34], MathVision [50] assess the mathematical reasoning ability in visual contexts. Differently, vision-centric benchmarks like MMVP [48], RealWorldQA [53], CV-Bench [47] curate questions that can be answered correctly *only* in the presence of the corresponding visual input. These vision-centric benchmarks ensure that the LVLM relies on its multimodal understanding capability for responding rather than the 'world knowledge' learned by the LLM. In particular, MMVP employs "CLIP-blind pair" images that CLIP [43] struggles to encode properly. While RealWorldQA evaluates the basic real-world spatial understanding capabilities of vision-language models, CV-Bench repurposes standard vision benchmarks like COCO [30], ADE20K [59], Omni3D [10] to assess their performance in classic vision tasks (such as spatial relationship, object count, depth order) within a multimodal context.

In contrast to the aforementioned benchmarks that test the vision-language models on high-level reasoning tasks, our SalBench strives to evaluate them on their low-level saliency perception capability. SalBench is a vision-centric benchmark, which repurposes the P3/O3 dataset [24] with language instructions to assess the LVLM ability to detect odd patterns in images that are visually salient and obvious to the human eye. Such an evaluation quantifies the vision-language model's ability to identify the distinctive target among numerous similar distractors, in terms of color, orinetation, shape, size, *etc*. Consequently, our SalBench is designed with an aim to measure the vision-language models' alignment with the human visual attention mechanism. Next, we describe our SalBench in detail.

3. The Saliency Benchmark: SalBench

We propose a saliency benchmark that assesses the capacity of foundational vision-language models to detect odd patterns that are salient for most humans. To achieve this, we augment the publicly available P3/O3 datasets [24] with language instructions. The P3 dataset comprises 2514 images (810 color, 864 orientation, and 840 size search arrays) arranged in a 7×7 grid [6, 52]. Each image contains distractors and a target, with pixel jitter applied to pre-

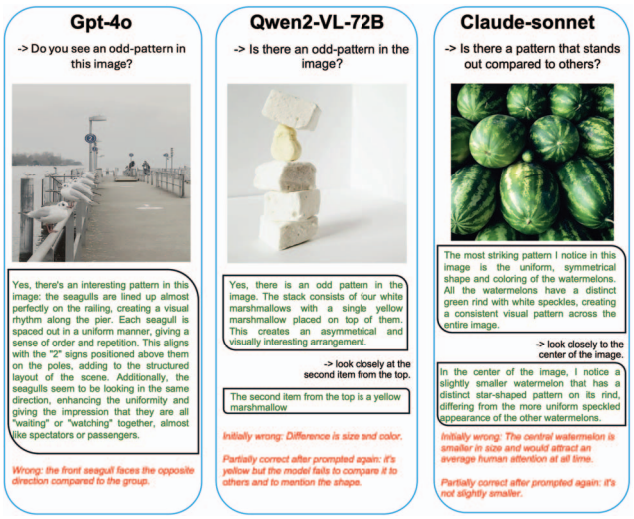

Figure 2. **Example responses from state-of-the-art vision-language models.** The model responses are in green, enclosed in boxes, while failure reasons are in red. These results highlight that the models often fail to recognize the prominent and salient features, which would naturally capture the attention of an average human observer. Best viewed zoomed-in.

vent perceptual grouping. On the other hand, the O3 dataset consists of 2001 real-world images featuring multiple similar objects (distractors) and a distinctive singleton (target). The target typically belongs to the same general category as the distractors but stands out in one or more feature dimensions (e.g., color, shape, size). O3 encompasses nearly 400 common object types, strongly emphasizing color singletons (37% differ by color alone, 47% by color and additional features). Other prominent distinguishing features include texture (33%), shape (26%), size (19%), and orientation (8%). The dataset provides a range of complexity, with distractor counts varying from 2 to over 50, offering a diverse challenge for saliency detection algorithms [24].

3.1. Task Creation

As discussed previously, in order to enhance the utility of the P3/O3 datasets as a benchmark for vision-language models, we propose to augment them with language instructions. Prompting the LVLM directly to predict the odd-pattern often results in incorrect responses from state-of-the-art LVLM. Few example responses for such a direct prompting are illustrated in Figure 2. The example conversations show that additional prompting or aid is required for the models to improve their responses. Consequently, to ensure better understanding of the model's perceptual capabilities, we create three novel tasks: Odd-One-Out Detection, Referring Odd-One-Out, and Visual Referring Odd-One-Out, as illustrated in Figure 1. We describe each task as follows:

(*i*) Odd-One-Out Detection: The basic task in this benchmark is *Odd-One-Out Detection*. In this task, given an input image, the model must identify which salient pattern from a predefined list of classes is present. For natural images, the list of features comprises: orientation, color, focus, shape, size, location, and pattern. This set of seven classes reflects the complexity of real-world scenes, where multiple factors such as size, shape, and location can combine to make an object stand out. The diversity of these classes captures saliency in natural images, making the task more challenging for models due to the potential interplay of various salient features. On the other hand, for synthetically generated images, the task is constrained to three patterns: color, shape, and orientation. This limitation is intentional to maintain control over the variables influencing saliency. By restricting the odd-pattern to a unique basic class rather than a combination of features, we can ensure that the model's focus is on a specific, isolated pattern. This simplification facilitates a clear assessment of the model's ability to detect saliency based on individual features.

(*ii*) Referring Odd-One-Out: This task builds on *Odd-One-Out Detection* by incorporating spatial information in the language prompt. The model receives an image and bounding box coordinates indicating the odd object's location via text. We test the model's ability to integrate textual and visual information, requiring it to focus on a specific image region using the language instruction and determine how the object there differs from the rest.

(*iii*) Visual Referring Odd-One-Out: This is a variant of the *Referring Odd-One-Out* task. Instead of using bounding box coordinates in text, this task visually highlights the odd object, typically with a red box. The model must identify the distinguishing features of this highlighted object compared to the other objects in the scene. This task evaluates the model's visual attention capabilities, requiring it to focus on a visually emphasized region-of-interest and determine the salient pattern in which the highlighted object stands out from its surroundings.

4. Evaluation

This section presents the evaluation results of the various models on SalBench. The analysis of model performances across zero-shot and few-shot learning settings reveals key insights into their capabilities and limitations.

Baselines: In our experiments, we evaluate a selection of state-of-the-art models spanning a range of parameter scales to benchmark the performance across model size variations. The lineup includes PaliGemma [7], Phi3, and Phi3.5 [1]; LLava 1.6 [31]; Idefics2 and Idefics3 [25]; VILA [29]; MiniCPM [20]; Qwen2-VL [51]; Molmo [14]; InternVL2 [46]; Llama3.2-Vision [38]; NVLM [13]; Claude [5]; and GPT-4o [2]. These models are considered the best LVLM according to the existing vision-language benchmarks. By encompassing a wide range of parameter

counts, from smaller and more efficient configurations to large-scale models, we can systematically analyze the impact of model size on task performance.

Setting: The salient targets can be different in one or more patterns, we devise the task as multi-label classification and report the accuracy of exact matches and the average F1 score over all categories. We evaluate model performance under both zero-shot and few-shot settings. For models that support multiple image inputs, we specifically conduct few-shot evaluation with 3-shots and 5-shots settings. While the zero-shot evaluation assesses the model's ability to generalize without exposure to specific examples, we expect that such a task isn't common and is likely to pose issues to most models. Hence, we also employ the few-shot evaluation to examine how the models leverage examples from the benchmark itself to solve the task. The default prompts for zero-shot evaluation of the three tasks are as described below:

(***i***) *Detection*: Context: Given this list of low-level visual features defined according to feature integration theory: Orientation, Color, Focus, Shape, Size, Location, Pattern. Task: Examine the provided image and identify the feature(s) in which one object notably differs from the others. Write out all applicable features separated by comma.

(***ii***) *Referring*: Context: This image depicts a scene with {num distractor} {object category}. Among those, one object category at location given by this bounding box $(x_{min}, y_{min}, x_{max}, y_{max})$, is different from the others. Given this list of low-level visual features defined according to feature integration theory: Orientation, Color, Focus, Shape, Size, Location, Pattern. Task: In which way does the special object differ from the rest. Write out all applicable features separated by comma.

(***iii***) *Visual Referring*: Context: This image depicts a scene with {num distractor} {object category}. Among those, one {object category} highlighted in a red box is different from the others. Task: Given this list of low-level visual features defined according to feature integration theory: Orientation, Color, Focus, Shape, Size, Location, Pattern. In which way does the special object differ from the rest. Write out all applicable feature(s) separated by comma.

Additional details of the few-shot prompts are provided in the supplementary material.

4.1. Results

In Table 1, we analyze the performance of various large vision-language models (LVLM) on our saliency benchmark, focusing on both synthetic and natural images splits (coming from P3 and O3, respectively). The synthetic and natural splits are denoted by SYN and NAT, respectively in the table. The tasks evaluated are detection, referring, and visual referring, measured using matching accuracy and F1 score metrics. Only F1 score metric is reported in Table 1 for clarity. The performance comparison with matching accuracy metric is provided in the supplementary material.

Method	Shot	Detection		Referring		Visual Referring	
		NAT	SYN	NAT	SYN	NAT	SYN
Claude-sonnet	0	48.2	86.7	51.1	90.3	53.9	87.7
NVLM-D-72B [13]	0	41.5	77.5	42.0	73.7	37.3	51.7
Molmo-7B[14]	0	32.0	67.2	32.4	38.0	33.1	28.4
Molmo-72B[14]	0	40.6	83.3	41.2	65.4	36.7	73.6
Llama3.2-Vision-11B [15]	0	32.1	48.7	29.1	52.4	29.7	52.4
PaliGemma-3B-448[7]	0	27.6	42.0	1.2	9.5	2.3	4.8
Phi3-4B[1]	0	32.1	41.2	32.8	55.3	32.8	47.2
	3	34.1	33.5	32.0	27.1	32.1	38.5
	5	31.1	17.0	32.1	18.9	32.2	46.7
Phi3.5-Vision-3.5B[1]	0	23.2	35.0	27.5	53.7	27.5	63.5
	3	23.3	19.5	28.8	41.0	28.8	20.8
	5	25.2	29.3	30.8	11.1	30.8	19.0
LLava 1.6-7B[31]	0	24.5	16.3	21.4	10.1	20.8	16.6
	3	7.0	16.4	15.2	8.8	17.8	17.0
	5	11.4	16.4	10.9	9.1	9.7	17.0
Idefics2-8B[26]	0	19.5	64.3	29.6	36.6	33.8	49.5
	3	21.1	66.3	28.4	34.2	31.1	39.6
	5	34.7	67.2	28.3	42.6	30.9	34.5
Idefics3-8B[25]	0	24.3	28.4	24.3	52.8	22.1	19.2
	3	26.9	40.3	26.9	20.67	21.9	40.6
	5	22.3	21.4	22.3	18.1	20.9	58.3
VILA-1.5-8B[29]	0	23.5	40.0	13.0	23.7	15.8	17.0
	3	25.1	17.0	28.8	21.2	28.8	17.0
	5	23.2	17.0	30.8	20.7	30.8	17.0
Qwen2-VL-1.5B[51]	0	19.2	26.3	22.1	20.6	20.9	20.2
	3	25.2	23.3	21.4	21.8	20.2	16.3
	5	25.3	23.8	21.7	16.5	20.9	17.7
Qwen2-VL-7B[51]	0	32.5	55.7	32.5	34.2	35.2	57.4
	3	35.6	53.8	36.0	17.0	34.1	64.2
	5	37.2	54.9	37.2	17.7	29.3	72.0
Qwen2-VL-72B[51]	0	41.6	88.8	44.6	93.6	41.7	74.7
	3	43.9	89.3	43.6	93.1	43.2	85.9
	5	43.9	89.9	44.9	92.6	42.3	87.9
InternVL-4B[12]	0	26.6	41.5	29.8	63.4	30.7	52.2
	3	27.7	17.0	27.4	25.3	29.5	41.7
	5	33.4	17.0	28.1	39.1	30.4	52.5
InternVL-2-8B[12]	0	20.0	58.7	23.0	71.9	24.8	23.0
	3	30.5	52.3	24.2	51.7	31.7	64.4
	5	27.8	43.9	25.0	53.7	31.4	50.5
GPT-4o	0	47.6	89.2	47.3	88.4	42.6	73.5
	3	38.9	88.4	37.5	87.7	35.7	86.7
	5	41.9	86.0	39.8	89.1	38.4	87.4

Table 1. **Performance comparison of vision-language models on the three tasks of SalBench for natural and synthetic image splits.** Natural and synthetic splits are denoted by NAT and SYN, respectively. The tasks are also evaluated under zero-shot and few-shot settings. The performance is reported in terms of F1 scores. We observe that performance of all models on these low-level perceptual tasks are lower in comparison to the standard vision-language benchmarks that test the models on high-level complex tasks. This shows that our SalBench offers another dimension for comprehensively evaluating LVLM, that is not present in the existing benchmarks in the literature.

Significant Performance Drops: All models exhibit a clear drop in performance on saliency tasks compared to their typical results on standard vision-language bench-

marks. This is evident in both matching accuracy and F1 scores across all tasks. For example, Qwen2-VL-72B achieves high scores on the synthetic split up to 89.9% F1 in the Detection task at the 5-shot setting. However, on the natural images split, its performance decreases significantly, achieving only 43.9% F1 in the Detection task at 5-shot. This indicates that models struggle with saliency tasks, especially in complex, real-world scenarios.

Better Performance on Synthetic Data: Models generally perform better on the synthetic images split than on the natural split. For instance, GPT-4o obtains an F1 score of 70.9% in the Detection task at 5-shot on synthetic split. In contrast, on the natural images split, GPT-4o achieves only 47.6% F1 in the same task and shot setting. This significant performance gap, often between 30% and 40%, suggests that models find it easier to process simplified synthetic images, where only one visual attribute varies, compared to the complexity of multi-label real-world natural images.

Model Size Influence: Larger models tend to outperform smaller ones on saliency tasks. In particular, Qwen2-VL-72B consistently achieves higher scores than its smaller variants, Qwen2-VL-7B and Qwen2-VL-1.5B. For the detection task on the the synthetic split, Qwen2-VL-72B reaches 89.9% F1 score in the detection task at 5-shot, while Qwen2-VL-7B achieves 54.9% and Qwen2-VL-1.5B only 23.8%. On the natural split, Qwen2-VL-72B attains 43.9% F1 score in detection at 5-shot, compared to 37.2% for Qwen2-VL-7B and 25.3% for Qwen2-VL-1.5B. This trend indicates that increased model capacity allows for better capture of complex patterns necessary for saliency tasks, a behavior not always observed in standard benchmarks. Similarly, GPT-4o shows strong performance, particularly in the Visual Referring tasks, indicating that larger models contribute to better saliency detection and reasoning capabilities over images. This trend suggests that increased model capacity allows for the capture of more complex patterns and relationships necessary for these tasks. This behavior is usually not seen in standard benchmarks such as VQAv2, as we see a marginal performance gap between large and small models.

Limited Impact of Few-Shot Learning: Increasing the number of shots does not consistently improve performance across models and tasks. Some models show comparable or even decreased performance with more shots. For example, Phi3.5-Vsion-3.5 on the synthetic split in the Visual Referring task has a F1 score of 63.5% at 0-shot, drops to 19.0 % at 5-shot. On the natural images split, GPT-4o performance in the Detection task decreases from 47.6% F1 score at 0-shot to 41.9% at 5-shot. This suggests that few-shot learning does not uniformly enhance performance on saliency tasks and that models may not effectively leverage additional examples in this context. The inability to generalize from limited examples suggests a lack in knowledge about saliency and visual attention for LVLM being worse when attempting to reason about the image.

Variability Across Tasks: The performance across tasks and splits for different models varies according to their visual grounding capability. For instance, Idefics3-8B achieves 21.4% F1 accuracy in Detection at 5-shot on the synthetic split but increases to 58.3% in Visual Referring. However, on the natural images split, the same model attains 22.3% in Detection at 5-shot but decreases to 20.9% in Visual Referring. This shows that it is likely easier for the model to perform visual grounding in synthetic images, compared to natural images. Differently, Phi3-4B shows improvements for Visual Referring, in comparison to Detection task on both splits, indicating better visual grounding capabilities. Furthermore, models with lower visual grounding capability often have higher performance for Detection task, compared to Visual Referring.

Challenges with Real-World Images: The performance drop from synthetic to real-world images is significant across models. Qwen2-VL-72B's score in the Detection task decreases from 89.9% on synthetic images split to 43.9% on natural split at 5-shot. GPT-4o shows a similar decline, from 86.0% on synthetic split to 41.9% on natural split. This suggests that models have difficulty handling the complexity and variability of real-world images, where multiple attributes and contextual factors are involved.

Inconsistent Few-Shot Performance Gains: Few-shot learning yields inconsistent and unclear performance gains. While some models show slight improvements, others do not benefit or even perform worse with additional shots. For example, GPT-4o on the natural images split achieves 47.6% F1 at 0-shot in Detection, drops to 38.9% at 3-shot, and then increases to 41.9% at 5-shot. Qwen2-VL-72B remains relatively stable across shot settings on the synthetic split, with F1 scores around 89% in Detection, while for the Visual Referring task, the performance improves by 13.2%. This inconsistency indicates that current models might not effectively utilize few-shot example contexts in saliency tasks.

4.2. Knowledge Testing of the Backbones

To identify where the shortcoming of vision-language models comes from in the saliency setting, we devise a test for individual components: LMMs and vision backbones.

Could LLMs solve FIT exams? We evaluated several Large Language Models (LLMs) on their understanding of Feature Integration Theory (FIT) [49] using a test comprising 50 questions collected from the internet. The larger models demonstrated good knowledge in this area. GPT-4o achieved the highest score with 97.5%, closely followed by Qwen2-72B-Instruct at 95.0%. Among the smaller models, Llama3-8B-Instruct attained a score of 82.5%. The Vicuna models achieved a score of 67.5%. These results indicate

Models	SYN		NAT	
	Top1	Top3	Top1	Top3
Random	24.6	53.2	14.6	36.3
Siglip-so400m-patch14-384	55.3	87.9	0.0	57.2
CLIP-ViT-Base-Patch32	42.1	71.7	12.0	59.8
CLIP-ViT-Base-Patch16	41.6	78.4	0.9	40.8
CLIP-ViT-Large-Patch14	41.2	78.6	0.0	31.6
CLIP-ViT-Large-Patch14-336	47.5	79.2	0.0	36.3

Table 2. **Zero-shot retrieval accuracy for natural and synthetic images using various vision backbones**. While the retrieval scores are reasonable for the synthetic images, the lower scores for natural images indicate the need for a stronger vision backbone in LVLM to capture low-level saliency information.

that even smaller LLMs possess an understanding of the visual stimuli concepts covered in FIT.

Visual representations retrieval: We investigate whether the feature output from the vision backbone are descriptive and discriminative enough about the saliency task. We compute the cosine similarity between the image embedding and the seven sentences for natural images and three sentences for the synthetic images, with embeddings of the form: "One object is different in {category} compared to the others". We take the top-k highest cosine similarity and count it as a match if any category in the top-k is present in the ground truth. The accuracy is reported in Table 2. Unlike knowledge test from LLM, we see clear gap in performance of vision model. While all models scores above random baseline, they scores relatively low, suggesting that in some cases, the information provided by the vision encoder are not discriminative enough for such low-level task. This highlights the need for stronger vision backbones for LVLM, capturing low-level spatial saliency information.

5. Training on saliency data

Data Generation: To construct the synthetic dataset, we employed web icons arranged in grid formations and systematically introduced controlled perturbations by altering the color, orientation, or size of randomly selected objects within each grid. This methodology yielded one million image-caption pairs for the alignment stage. For the supervised fine-tuning stage, we synthesized additional images with corresponding questions and answers based on the introduced odd-patterns. The questions were designed to be either open-ended, enhancing the models' saliency understanding capabilities, or multiple-choice, aligning with the benchmark format.

Training: We selected LLama3.1 8B [15] and Qwen2-7B [54] as the LLMs. For the vision encoders, we incorporated both CLIP [43] and SigLip [58], resulting in four model variants. We followed a Llava like training recipe [31]. During the alignment stage, we use one million

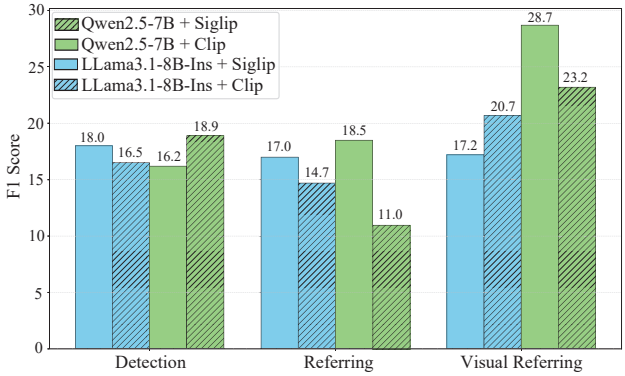

Figure 3. **Performance comparison of trained LVLM on saliency generated data**, using two LLMs (Qwen2.5-7B and LLama3.1-8B) paired with two vision encoders (Siglip and Clip) across three tasks: Detection, Referring, and Visual Referring. The F1 scores suggest that training on such data does not help improve the performance on SalBench.

saliency data with two million image-caption pairs sourced from natural image datasets to promote robust multimodal representation learning. In the instruction-tuning stage, we integrated two million data points from the recently introduced Cambrian dataset [47] with the generated one million instruction saliency data. We employed the same hyperparameters and training configurations as specified in Llava [31] to ensure consistency and facilitate fair comparison of results.

Performance: As shown in Table 3, despite training on in-distribution saliency data, the performance of all model variants across the P3 tasks remains low. Detection scores are around 16–19%, with LLama3.1-8B paired with CLIP achieving the highest at 18.9%. For the referring task, scores are even lower and more variable, ranging from 11.0% to 18.5%. The visual reference task shows slightly better results, particularly for LLama3.1-8B with SigLip, which attains a score of 28.7%, yet this is still insufficient for practical applications. These consistently low scores indicate that the models struggle to capture the salient features within the images, even when trained on data specifically designed to highlight these aspects. The lack of improvement suggests that the current architectures and training recipe are not suitable for capturing the of saliency information in images.

6. Analysis

In this section, we study the factors contributing to the poor performance of LVLM on the visual attention benchmark.

6.1. Accuracy Breakdown by Task and Difficulty

Bias towards color: As seen in Table 3, among the low-level features what were asked to identify, color is the only

Model	Category	Level	Accuracy		
			Detection	Referring	Visual Referring
Qwen2-VL-72B	Orientation	Easy	98.6	97.1	88.3
		Medium	95.8	96.8	83.0
		Hard	95.7	96.2	71.7
GPT-4o	Orientation	Easy	96.2	96.5	97.6
		Medium	96.2	89.2	94.7
		Hard	98.6	88.5	96.2
Qwen2-VL-72B	Size	Easy	94.2	98.8	62.3
		Medium	50.5	74.3	31.9
		Hard	46.0	62.0	33.3
GPT-4o	Size	Easy	93.3	96.9	42.5
		Medium	59.1	59.1	39.1
		Hard	36.8	35.3	11.3
Qwen2-VL-72B	Color	Easy	100.0	100.0	99.1
		Medium	100.0	100.0	98.6
		Hard	60.1	80.2	73.5
GPT-4o	Color	Easy	99.8	99.0	96.1
		Medium	100.0	100.0	93.3
		Hard	66.1	66.2	57.9

Table 3. **Accuracy across different difficulty levels on synthetic images.** It can be seen that orientation and color are robust over the three levels, whereas size scores drop for both models.

# Distractors	Orientation	Color	Focus	Shape	Size	Location	Pattern	Avg F1
Qwen2-VL-72B								
<7	18.4	92.8	18.6	53.4	39.7	30.3	57.9	44.5
7-15	9.3	93.7	13.0	55.8	37.8	31.6	41.2	40.4
15-25	23.5	95.8	11.1	59.1	39.5	0.0	37.5	38.1
>25	16.7	96.1	15.0	52.2	36.5	20.0	25.4	37.4
GPT-4o								
<7	37.1	88.2	38.1	49.4	35.7	23.6	57.0	47.0
7-15	46.2	90.1	40.0	47.5	32.6	27.5	53.3	48.2
15-25	43.8	91.5	45.2	47.5	31.0	12.5	45.1	45.2
>25	40.5	89.8	48.4	48.1	30.4	4.8	29.4	41.6

Table 4. **Accuracy for varying the number of distractors for real images.** The performance linearly decreases when adding more distractors to the scene.

features directly provided by the data in the form of RGB images. Other features such as size, and shape need to be encoded in higher level representation. We break down the scores in supplementary by category, and observe that all models perform higher when the target class is color.

On the synthetic images (originating from P3 dataset) of the SalBench, we assess two best models (Qwen2-VL-72B and GPT-4o) across the categories at different levels. The difficulty levels within each category are defined based on specific attributes of the target object relative to distractors:

- **Orientation:** Difficulty is determined by the angular difference in rotation between the target object and the distractors. *Hard*: Rotation differences ranging from 0 to 30 degrees, where minimal rotational disparity renders visual cues less distinguishable. *Medium*: Rotation differences between 30 to 60 degrees. *Easy*: Rotation differences from 60 to 90 degrees, facilitating easier identification due to clear angular differences.
- **Size:** Levels are determined based on the area ratio between the target object and the distractors, calculated using their heights and widths. *Hard*: Ratio between 0.5 and 1.5, where the target and distractors are close in size. *Medium*: Ratio between 0.3 and 0.5 (target smaller than distractors) or between 1.5 and 3.0 (target larger than distractors), representing moderate size differences. *Easy*: Ratio less than 0.3 or greater than 3.0, indicating that the target is much smaller or much larger than the distractors, making it easily distinguishable.
- **Color:** Difficulty is assessed by calculating the euclidean distance between the RGB color values of the target and distractors. The RGB values are extracted from hexadecimal color codes, to compute the color distance. *Hard*: Color Distance less than or equal to 50, indicating high color similarity and challenging discrimination. *Medium*: Color Distance between 50 and 100, indicating moderate color differences. *Easy*: Color Distance greater than 100, where the target's color clearly contrasts with that of the distractors, making identification easier.

Results and Analysis: The performance metrics for both models are detailed in Table 3. The following observations highlight the models' capabilities and limitations across difficulty levels:

Orientation: Qwen2-VL-72B and GPT-4o demonstrate high accuracy in the Detection and Referring tasks across all difficulty levels, indicating robust performance in recognizing and describing objects with rotational variations. Specifically, Qwen2-VL-72B maintains Detection accuracy above 95% even at the hard level. However, in the Visual Referring task, Qwen2-VL-72B exhibits a decline from 88.3% accuracy at the easy level to 71.7% at the hard level. In contrast,GPT-4o sustains consistently high Visual Referring accuracy, exceeding 94% across all difficulty levels.

Size: The models' performance deteriorates as the difficulty increases. For Qwen2-VL-72B, Detection accuracy drops from 94.2% at the easy level to 46.0% at the hard level. Similarly, Visual Referring accuracy drops from 62.3% to 33.3%. GPT-4o follows a comparable trend, with Detection accuracy decreasing from 93.3% to 36.8% and Visual Referring accuracy from 42.5% to 11.3%. This suggest both models struggle when size disparities are minimal, highlighting a limitation in perceiving subtle scale differences.

Color: Both models achieve near-perfect accuracies at the easy and medium levels. Qwen2-VL-72B attains 100% accuracy in Detection and Referring tasks, while GPT-4o records accuracies exceeding 99%. However, at the hard level—characterized by minimal color differences—there is a notable decline. Qwen2-VL-72B's Detection accuracy decreases to 60.1%, and GPT-4o's to 66.1%.

Figure 4. **Perception change test.** GPT-4o fails to identify that the individuals talking to the old man were switched.

6.2. Impact of Distractor Quantity on Detection

We study how the number of distractors in an image affects the performance of the model, on the natural images Detection task. The aim is to understand the models' robustness to increased visual complexity. As shown in Table 4, both models exhibit a decline in average F1 scores as the number of distractors increases. For Qwen2-VL-72B, the average F1 score decreases from 44.5% when there are fewer than 7 distractors to 37.4% when there are more than 25 distractors. GPT-4o shows a similar trend, with its average F1 score dropping from 47.0% to 41.6% over the same range. AlthoughGPT-4o consistently outperforms Qwen2-VL-72B across most distractor quantities, the performance gap narrows as the number of distractors increases. This convergence indicates that both models are similarly challenged by increased visual complexity.

In fact, both models maintain high accuracy in the *Color* category regardless of distractor quantity. However, performance declines are evident in categories requiring spatial reasoning and attention to specific object attributes. In the *Location* category, Qwen2-VL-72B's F1 score drops dramatically from 30.3% with fewer than 7 distractors to 20.0% with more than 25 distractors. Interestingly, in the **Shape** category, both models maintain relatively stable performance across different distractor quantities. Overall, these findings highlight that while certain visual features like color and, to some extent, shape remain reliable even, increased numbers of distractors hurt the performance.

6.3. Why detecting saliency is important?

While SalBench might not contribute at measuring LVLM explicit ability for downstream applications [11, 17, 27, 34, 55, 57]. Detecting visual saliency is a fundamental aspect of human perception, for vision-language models replicating this capability is critical for several reasons:

Applications in Robotics and Autonomous Agents: In robotics, the ability to detect salient features is crucial for navigation, object recognition, and interaction within dynamic environments. Recently introduced works leverage LVLM for such applications [9, 23]. As shown in Figure 4, the individual holding the paper was switched, andGPT-4o did not detect this change, we better have this ability for robots with LVLM as engines not to miss salient events. Incorporating the saliency detection capabilities could enable robots and agents to prioritize important stimuli, focus on objects of interest, and respond appropriately to unexpected changes, thereby enhancing their autonomy and operational efficiency. Hence, using SalBench could give insights about the robustness of LVLM for real world deployment.

LVLM agents: Detecting saliency is important in using these models as agents for website navigation. Designers use principles from FIT to make important elements, such as call-to-action buttons, stand out by contrasting their features (e.g., color, size, or shape). When LVLM are deployed as agents to interact with web environments [19, 41], effective saliency detection becomes needed for efficient navigation and interaction. For instance, an agent tasked with automating web browsing or performing tasks like form filling, data extraction, or content moderation must quickly identify and focus on salient elements such as buttons, links, advertisements, or notifications. By accurately detecting visually prominent features, the agent can operate faster and more effectively, closely mimicking human browsing behavior and enhancing overall performance.

7. Conclusion

We introduced a saliency benchmark, called SalBench, aimed at evaluating the abilities of large vision-language models (LVLM) in detecting low-level visually-salient features such as color, orientation, and size, which are the building blocks of the human's visual cortex. Through three tasks, Odd-One-Out Detection, Referring Odd-One-Out, and Visual Referring Odd-One-Out, we assessed how well LVLM align with human visual attention mechanisms using both synthetic and natural images. Our evaluation revealed clear limitations in current LVLM, with even advanced models like GPT-4o achieving only 46% accuracy on simple saliency detection tasks. This highlights a clear gap between LVLM and human perceptual capabilities in processing fundamental visual features. SalBench serves as a tool for benchmarking and improving the perceptual alignment of LVLM with human attention mechanisms. Enhancing these models' low-level perceptual abilities is essential for advancing toward more human-like visual understanding of images.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 3, 4

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[3] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 2

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[5] Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. 3

[6] SP Arun. Turning visual search time on its head. *Vision research*, 74:86–92, 2012. 2

[7] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 3, 4

[8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. 2

[9] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 8

[10] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13154–13164, 2023. 2

[11] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1, 2, 8

[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. 1, 4

[13] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 3, 4

[14] 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. 3, 4

[15] 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. 4, 6

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2

[17] 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, 2, 8

[18] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 2

[19] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290, 2024. 8

[20] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. 2024. *URL https://doi.org/10.48550/arXiv*, 2404. 3

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[22] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 2

[23] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 8

[24] Iuliia Kotseruba, Calden Wloka, Amir Rasouli, and John K Tsotsos. Do saliency models detect odd-one-out targets? new datasets and evaluations. *arXiv preprint arXiv:2005.06583*, 2020. 2, 3

[25] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 3, 4

[26] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 4

[27] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 8

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 2

[29] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 3, 4

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3, 4, 6

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1, 2

[33] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[34] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv e-prints*, 2023. 1, 2, 8

[35] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2

[36] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021. 2

[37] 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*, 2022. 2

[38] Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable modelsy. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024. 3

[39] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 2

[40] Hans Moravec. *Mind children: The future of robot and human intelligence*. Harvard University Press, 1988. 1

[41] Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang.

Screenagent: A vision language model-driven computer control agent. *arXiv preprint arXiv:2402.07945*, 2024. 8

[42] OpenAI. Gpt-4o. https://platform.openai.com/docs/models#gpt-4o, 2024. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6

[44] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 1, 2

[45] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 2

[46] InternVL team. Internvl2: Better than the best — expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. https://internvl.github.io/blog/2024-07-02-InternVL-2.0, 2024. 3

[47] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 6

[48] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2

[49] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2, 5

[50] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset, 2024. 2

[51] 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. 3, 4

[52] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature human behaviour*, 1(3):0058, 2017. 2

[53] xAI. Grok-1.5v. https://x.ai/blog/grok-1.5v, 2024. 2

[54] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 6

[55] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 1, 8

[56] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 2

[57] 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, 8

[58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 6

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 2

[60] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1