Visual Intelligence

RESEARCH Open Access

Effectiveness assessment of recent large vision-language models

Yao Jiang[1,2†], Xinyu Yan[1,3†], Ge-Peng Ji[5], Keren Fu[2*], Meijun Sun[3], Huan Xiong[1,4*], Deng-Ping Fan[6] and Fahad Shahbaz Khan[1]

Abstract

The advent of large vision-language models (LVLMs) represents a remarkable advance in the quest for artificial general intelligence. However, the models' effectiveness in both specialized and general tasks warrants further investigation. This paper endeavors to evaluate the competency of popular LVLMs in specialized and general tasks, respectively, aiming to offer a comprehensive understanding of these novel models. To gauge their effectiveness in specialized tasks, we employ six challenging tasks in three different application scenarios: natural, healthcare, and industrial. These six tasks include salient/camouflaged/transparent object detection, as well as polyp detection, skin lesion detection, and industrial anomaly detection. We examine the performance of three recent open-source LVLMs, including MiniGPT-v2, LLaVA-1.5, and Shikra, on both visual recognition and localization in these tasks. Moreover, we conduct empirical investigations utilizing the aforementioned LVLMs together with GPT-4V, assessing their multi-modal understanding capabilities in general tasks including object counting, absurd question answering, affordance reasoning, attribute recognition, and spatial relation reasoning. Our investigations reveal that these LVLMs demonstrate limited proficiency not only in specialized tasks but also in general tasks. We delve deep into this inadequacy and uncover several potential factors, including limited cognition in specialized tasks, object hallucination, text-to-image interference, and decreased robustness in complex problems. We hope that this study can provide useful insights for the future development of LVLMs, helping researchers improve LVLMs for both general and specialized applications.

Keywords: Large vision-language models (LVLMs), Recognition, Localization, Multi-modal understanding

1 Introduction

The emergence of large language models (LLMs) [1, 2] has sparked a revolution in the field of natural language processing, owing to their promising generalization and reasoning capabilities. Motivated by this progress, researchers have pioneered the development of powerful large vision-language models (LVLMs) [3–7], leveraging the impressive capabilities of LLMs to enhance comprehension of visual semantics. This advance particularly improves model performance in complex vision-language tasks [4, 8, 9], and represents a major step toward artificial general intelligence (AGI). AGI refers to intelligent systems that are capable of solving any task that can be performed by humans or animals [10]. Generally, tasks performed by humans can be divided into general and specialized tasks according to whether special domain knowledge is required. Therefore, the capabilities of LVLMs can be categorized into these two aspects accordingly, and both of them are essential for LVLMs on the path toward AGI.

Recently, many studies have assessed and investigated the general and specialized capabilities of LVLMs [8, 9, 11–15]. Qin et al. [9] conducted empirical studies encompassing various general tasks, such as object detection and counting to evaluate the visual understanding capabilities

*Correspondence: fkrsuper@scu.edu.cn; huan.xiong.math@gmail.com
[1]Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi 999041, UAE
[2]Sichuan University, Chengdu 610065, China
Full list of author information is available at the end of the article †Equal contributors

of Google Bard. Fu et al. [15] introduced a comprehensive evaluation benchmark to assess the perceptual and cognitive capabilities of recent LVLMs on general tasks (e.g., optical character recognition and object counting). Zhang et al. [11] explored the potential of GPT-4V [5] in visual anomaly detection, while Tang et al. [12] generalized Shikra [7] to challenging camouflaged object detection scenarios without training. However, as these studies primarily focus on evaluating the general capabilities of LVLMs [8, 9, 15] or exploring the effectiveness of a particular LVLM in a specialized domain [11–14], there is a lack of quantitative analysis regarding the performance of recent LVLMs in a diverse range of specialized tasks, leading to an insufficient understanding of their capabilities.

In this paper, we conduct a comprehensive assessment of several recent open-source LVLMs, spanning a diverse array of challenging specialized and general tasks. Our evaluation platform is illustrated in Fig. 1. To evaluate the ability of LVLMs to perform specialized tasks, we select three recent open-source LVLMs (MiniGPT-v2 [4], LLaVA-1.5 [6], and Shikra [7]) and conduct quantitative assessment on six challenging specialized tasks in three different application scenarios: natural, healthcare, and industrial. For natural scenarios, we select salient object detection (SOD) [17–19], transparent object detection (TOD) [20] and camouflaged object detection (COD) [21, 22], as these tasks involve targets that are increasingly rare in real life and have increasingly complex characteristics, thereby presenting distinct challenges to LVLMs. In the field of healthcare, the effectiveness of LVLMs is evaluated by skin lesion detection [23] and polyp detection [24], which show prominent and slightly weaker visual features, respectively. Besides, anomaly detection (AD) [25], a vital task in industrial scenarios, is also selected for assessment. In academia, these six tasks come with tailored datasets and cover broad specialized domains, thereby enabling comprehensive evaluation of specialized capabilities of LVLMs. As illustrated in Fig. 1, given inherent challenges posed by these tasks in terms of recognizing and localizing target objects, we employ tailored prompts to assess the recognition (Sect. 2) and localization (Sect. 3) capabilities of the models. Furthermore, we conduct empirical investigations on a universal dataset (COCO [16]) that is free from domain-specific expertise. We refrain from specifying particular object types ("camouflaged", "transparent", or other) in prompts, aiming to explore multimodal understanding capabilities (Sect. 4) of the above-mentioned models and GPT-4V in general tasks (i.e., object counting, absurd question answering, affordance reasoning, attribute recognition, and spatial relation reasoning). The assessed prominent LVLMs, include MiniGPT-v2 [4], LLaVA-1.5 [6], Shikra [7], and GPT-4V [5], all of

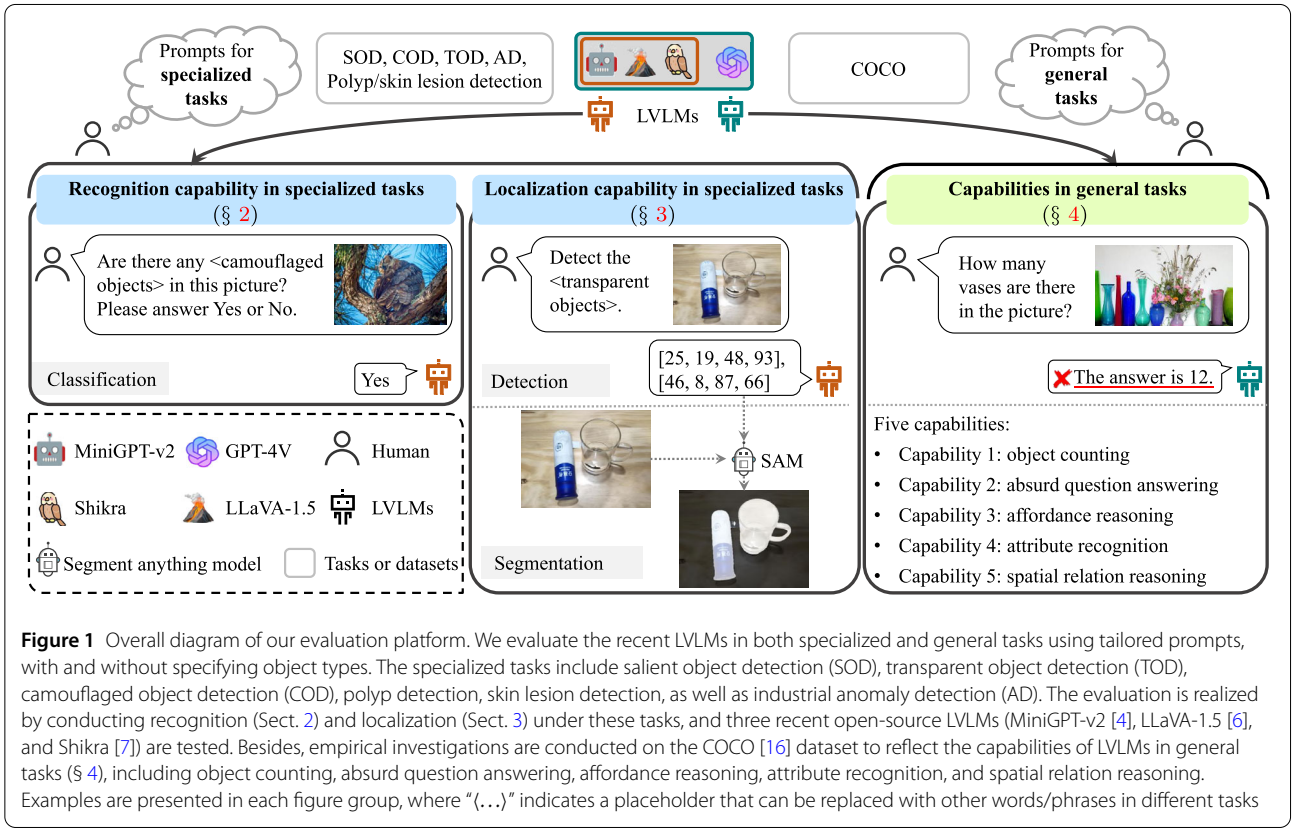

Figure 1 Overall diagram of our evaluation platform. We evaluate the recent LVLMs in both specialized and general tasks using tailored prompts, with and without specifying object types. The specialized tasks include salient object detection (SOD), transparent object detection (TOD), camouflaged object detection (COD), polyp detection, skin lesion detection, as well as industrial anomaly detection (AD). The evaluation is realized by conducting recognition (Sect. 2) and localization (Sect. 3) under these tasks, and three recent open-source LVLMs (MiniGPT-v2 [4], LLaVA-1.5 [6], and Shikra [7]) are tested. Besides, empirical investigations are conducted on the COCO [16] dataset to reflect the capabilities of LVLMs in general tasks (§ 4), including object counting, absurd question answering, affordance reasoning, attribute recognition, and spatial relation reasoning. Examples are presented in each figure group, where "⟨…⟩" indicates a placeholder that can be replaced with other words/phrases in different tasks

which have garnered significant research attention as key players in the field. Among them, three accessible open-source models, i.e., MiniGPT-v2, LLaVA-1.5, and Shikra, are selected to ensure feasibility and reproducibility of the evaluation in specialized tasks.

Our investigations reveal that while these models show strong potential for specialized tasks, they still exhibit sub-optimal performance and limited cognitive capabilities. This reveals their inadequate transfer ability in this particular context. Performance issues are further magnified by typical weaknesses of LVLMs such as object hallucination, text-to-image interference, and decreased robustness in complex problems. In addition to the shortcomings revealed in specialized tasks, these models also show significant room for improvement in general tasks, particularly in object counting, spatial reasoning, and absurd question answering.

In summary, the main contributions of this paper are three-fold: (1) We construct an evaluation platform comprising six specialized tasks and five general tasks to assess the effectiveness of LVLMs. (2) On the evaluation platform, we evaluate the specialized capabilities of three recent open-source LVLMs and also the general capabilities of four LVLMs. (3) We analyze their performance and limitations for both specialized and general tasks, and discuss the future development and application of LVLMs.

2 Recognition via LVLMs in specialized tasks

When LVLMs are applied in these specialized tasks, recognizing these target objects is a crucial step, which reflects models' global understanding of such tasks and directly influences their effectiveness. Therefore, we first conduct quantitative evaluation of their recognition capabilities on the aforementioned six specialized tasks. Subsequently, we carry out additional tests to delve into failure cases and gain further insights.

2.1 Quantitative investigation

2.1.1 Experimental setup

Recognition in specialized tasks involves determining the existence of targets and classifying them. The first evaluation of recognition capabilities is to judge object existence, requiring models to answer either "Yes" or "No" to questions such as "Are there any ⟨camouflaged objects⟩ in the picture? Please answer Yes or No.", as demonstrated in Fig. 1. The placeholder "⟨…⟩" in the queries denotes flexible words/phrases that can be substituted in different tasks, such as "polyps" in polyp detection. The evaluation considers two different setups: the full set, which includes both positive and negative samples, and the positive set, which includes only positive samples.

Beyond the first evaluation, we delve deeper into the fine-grained recognition ability of LVLMs by asking them to categorize targets. Our method is to prompt LVLMs to designate the most suitable category for a target object from a pre-defined set of potential categories (w/ vocabulary). Within this experiment, the questions such as "Which of the following is the most likely category for the camouflaged object in the picture? 'seahorse, mantis, spider…' " are used. The pre-defined set contains all categories that appear in the dataset. Besides, another evaluation is considered, featuring an open-vocabulary inquiry without giving a pre-defined set (w/o vocabulary). In this test, a straightforward question like "What is the camouflaged object in the picture?" is used.

The versions of LLavA-1.5 [6], Shikra [7], and MiniGPT-v2 [4] that are equipped with language models of approximately 7 billion parameters are selected for evaluation. All configurations of each model are set as default during evaluation. Since all tests in this paper are based on the above configurations, we will not mention again in the following sections.

2.1.2 Metrics

As for the first evaluation, accuracy ($\mathcal{A}$) is employed to measure the performance of LVLMs in judging object existence, while the probability of positive responses (responses indicating "yes") on the full set is also reported for reference. $\mathcal{A}$ and the probability of positive responses ($\mathcal{Y}$) can be formulated as follows:

$$\mathcal{A} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}, \tag{1}$$

$$\mathcal{Y} = \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}, \tag{2}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

For fine-grained recognition, LVLMs typically select categories from a pre-defined set when available, enabling direct matching with labels for accuracy assessment. However, in the absence of such a set, the generated categories exhibit significant variation, posing challenges in directly evaluating correctness through class matching. Hence, we utilize accuracy ($\mathcal{A}^*$) and semantic similarity ($\mathcal{S}$) [26] to measure the performance in these two settings, respectively. The former quantifies the fraction of responses that contain correct category names, while the latter quantifies the semantic similarity between responses and ground truth labels. Considering that LVLMs may occasionally generate similar categories not included in the pre-defined set, $\mathcal{S}$ is also employed to evaluate the performance of the w/ vocabulary setting.

2.1.3 Benchmark datasets

A total of 10 datasets from SOD (DUTS [27] and SOC [28]), COD (COD10K [21]), TOD (Trans10K [20]), polyp detection (ColonDB [24], ETIS [29], and CP-CHILD-B [30]),

skin lesion detection (ISIC [23]), and AD (MVTec AD [25] and VisA [31]) are employed to evaluate the performance of LVLMs in determining the existence of targets. Among these datasets, SOC, COD10K, CP-CHILD-B, MVTec AD, and VisA, which contain both positive and negative samples, are used to construct the full set, while the remaining datasets are utilized to form the positive set. The proportions of positive samples in SOC, COD10K, CP-CHILD-B, MVTec AD, and VisA are 50%, 50.7%, 25%, 72.9%, and 55.5%, respectively.

COD10K, the only dataset that provides category labels for each target, is utilized to evaluate the fine-grained recognition ability of LVLMs. Since judging target existence in negative samples is certainly challenging for LVLMs, we exclude the interference and use only the positive samples of COD10K to more accurately evaluate the fine-grained recognition ability of LVLMs.

2.1.4 Result analyses and discussions

Evaluation results of existence determination on the full set and positive set, and fine-grained recognition are detailed in Tables 1-3. The absence of negative samples leads to TN = 0 and FP = 0, and hence $\mathcal{A}$ in Table 2 is equivalent to $\mathcal{Y}$ in Table 1. Three observations from these results are as follows.

Over-positive issue From the results in Table 1 and the proportion of positive samples in each dataset (in Sect. 2.1.3), we can observe that these models consistently yield a greater proportion of positive responses ($\mathcal{Y}$) compared to the proportion of positive samples. Especially on SOC and CP-CHILD-B, these LVLMs generally achieve $\mathcal{Y}$ higher than 0.9, while the proportions of positive samples in these datasets are only 50% and 25%. This indicates that the models tend to give positive responses, which is further proved on the positive sets in Table 2, where extremely high scores on $\mathcal{A}$ (e.g., 1.000) are achieved (particularly for LLaVA-1.5). The reason behind this phenomenon could be that most of the samples learned by these LVLMS during the training are positive image-text pairs, which makes them over-positive and thus have a tendency to answer "yes" to the questions [32, 33].

Limited performance in determining existence Though notably high accuracy ($\mathcal{A}$) in Table 2 are achieved by LVLMs, the inclusion of negative samples results in an overall decrease in accuracy. As shown in Table 1, most accuracies drop below 0.7, indicating an inadequate recognition ability of LVLMs in determining the existence of targets, particularly in the case where negative samples are presented. Among these models, LLaVA-1.5 shows better

Table 1 Experimental results for three LVLMs regarding the presence of targets on the full sets. We present the probability of positive answers ($\mathcal{Y}$, representing the percentage of "yes"). The highest accuracy ($\mathcal{A}$) score is highlighted in bold

Model	Natural scenarios		Healthcare	Industrial scenarios	
	SOC [28]	COD10K [21]	CP-CHILD-B [30]	MVTec AD [25]	VisA [31]
	$\mathcal{A}/\mathcal{Y}$	$\mathcal{A}/\mathcal{Y}$	$\mathcal{A}/\mathcal{Y}$	$\mathcal{A}/\mathcal{Y}$	$\mathcal{A}/\mathcal{Y}$
MiniGPT-v2 [4]	0.513/0.987	0.580/0.909	0.250/0.990	0.695/0.874	0.543/0.875
LLaVA-1.5 [6]	**0.618**/0.883	**0.776**/0.427	0.268/0.983	**0.750**/0.979	0.562/0.993
Shikra [7]	0.528/0.973	0.535/0.053	**0.285**/0.945	0.728/0.562	**0.617**/0.251

Table 2 Experimental results for three LVLMs regarding the presence of targets on the positive sets. The highest accuracy score is marked in bold. Given the absence of negative samples in the positive set, resulting in $\mathrm{TN} = 0$ and $\mathrm{FP} = 0$, the metric $\mathcal{A}$ in this table is equivalent to $\mathcal{Y}$

Model	Natural scenarios		Healthcare		
	DUTS [27]	Trans10K [20]	ColonDB [24]	ETIS [29]	ISIC [23]
	$\mathcal{A}$	$\mathcal{A}$	$\mathcal{A}$	$\mathcal{A}$	$\mathcal{A}$
MiniGPT-v2 [4]	0.853	0.964	0.824	0.847	0.952
LLaVA-1.5 [6]	0.999	**1.000**	**1.000**	**0.985**	**1.000**
Shikra [7]	**1.000**	0.988	0.968	0.954	0.998

Table 3 Quantitative results of three LVLMs for classifying camouflaged objects. The best results are marked in bold. We conduct classification solely on the positive samples within COD10K [21]

Setting	Metric	MiniGPT-v2 [4]	Shikra [7]	LLaVA-1.5 [6]
w/ vocabulary	$\mathcal{A}^*$	0.285	0.154	**0.436**
w/ vocabulary	$\mathcal{S}$	0.567	0.545	**0.673**
w/o vocabulary	$\mathcal{S}$	0.607	0.608	**0.655**

recognition capabilities for camouflaged objects, achieving higher accuracy ($\mathcal{A}$) while obtaining $\mathcal{Y}$ scores that are close to the proportions of positive samples in COD10K. In contrast, Shikra shows extremely bad results (on $\mathcal{Y}$) on COD10K due to its frequent misclassification of positive samples, indicating its less capability in recognizing camouflaged objects.

Struggling with classifying camouflaged objects The results in Table 3 clearly demonstrate that these LVLMs struggle with classifying camouflaged objects. Although LLaVA-1.5 achieves the highest scores, its performance is still unsatisfactory. The unsatisfactory performance could be attributed to various factors. First, these models may face challenges in identifying camouflaged objects that closely resemble the background, as indicated by their unsatisfactory recognition accuracy in Table 1. Second, the category of camouflaged objects may lie beyond the models' domain of knowledge, hindering their capability to match objects with their categories accurately. Additionally, the extended length of the prompt, stemming from the incorporation of the pre-defined set, may impede the models' comprehension. This aligns with the results in Table 3, where MiniGPT-v2 and Shikra demonstrate improved performance ($\mathcal{S}$) when the pre-defined set is excluded (i.e. w/o vocabulary), as opposed to when the vocabulary is provided (i.e. w/ vocabulary).

2.2 Uncovering insights into failure cases

Recalling that these models encounter challenges in differentiating negative samples, so we conduct tests on representative negative samples to gain insight into the potential causes of this phenomenon. LVLMs are prompted to provide additional description or reasoning when determining the existence of targets. The results are illustrated in Fig. 2, where three potential factors are derived.

Limited cognition towards special object types As illustrated in the first example of Fig. 2, when presented with the question "Is there camouflaged object in the picture? What is it?", MiniGPT-v2 erroneously recognizes the "small black rock" as a camouflaged object, while LLaVA-1.5 misclassifies a "plant" as such. These models classify rocks and plants as camouflaged objects just because of their visual resemblance to the surroundings, indicating their limited knowledge of camouflage. This phenomenon

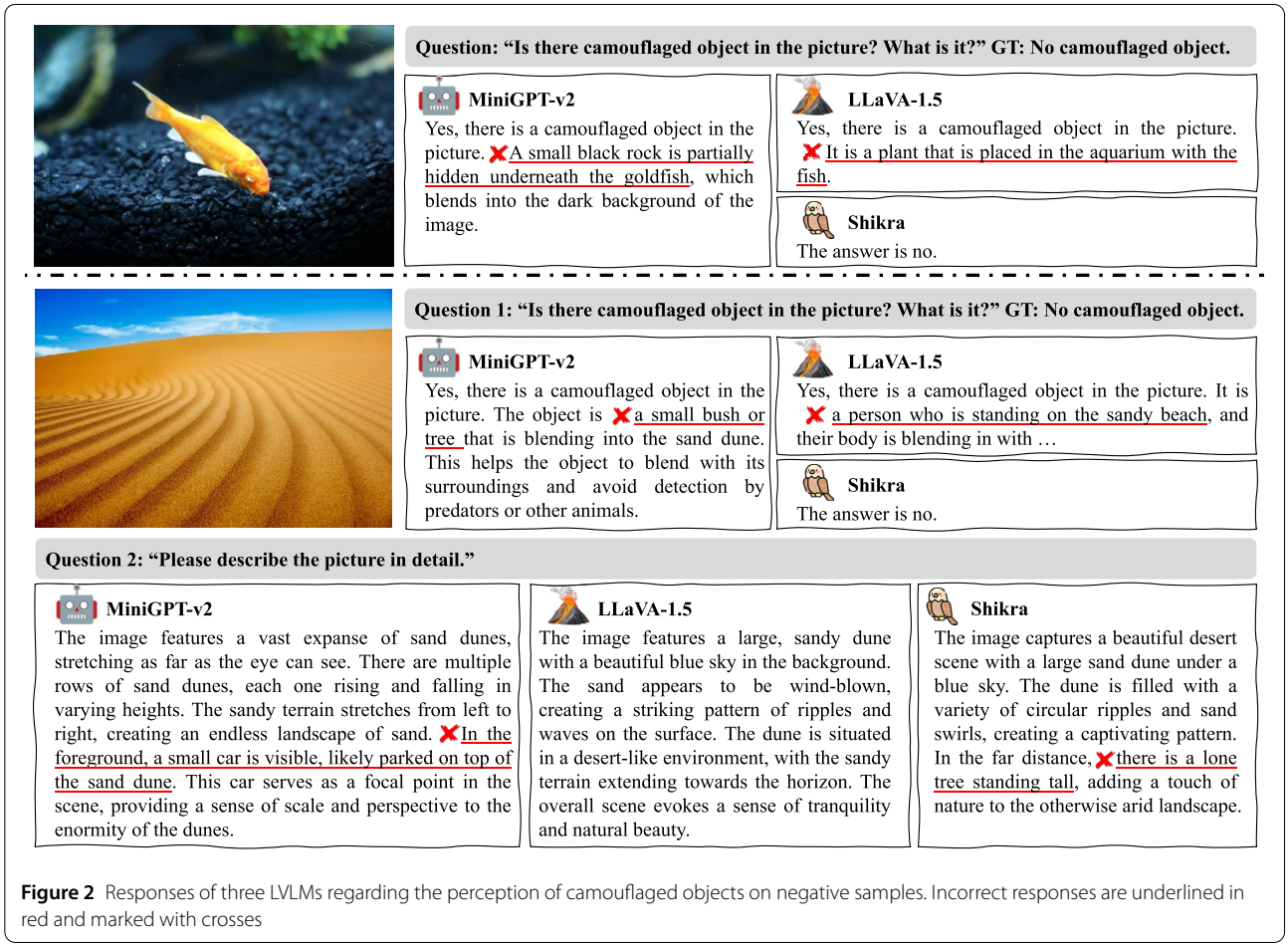

Figure 2 Responses of three LVLMs regarding the perception of camouflaged objects on negative samples. Incorrect responses are underlined in red and marked with crosses

also occurs in other specialized tasks, e.g., anomaly detection, implying their limited cognition on special object types.

Object hallucinations Object hallucination, which involves imagining objects in the response but not present in the image [32, 34], could impact the recognition capability of LVLMs in specialized tasks. For instance, as demonstrated by the answers to "Is there a camouflaged object in the picture? What is it?" in the second example of Fig. 2, LLaVA-1.5 states that "a person is standing on the sandy beach", while MiniGPT-v2 mentions the presence of "small bush or tree". These objects can interfere with target recognition [12], resulting in decreased recognition performance when determining object presence.

Text-to-image interference The inadequate performance in determining the presence of targets may also be attributed to text-to-image interference, which originates from the textual prompts supplied to the models [34]. As shown in the second example in Fig. 2, when prompted with "Please describe the picture in detail", LLaVA-1.5 provides an accurate description of the image. However, when prompted with "Is there a camouflaged object in the picture? What is it?", the mention of the "camouflaged object" in the prompt may interfere with the answers, resulting in hallucination and misjudgment of LLaVA-1.5.

2.3 Summary

Section 2 evaluates the recognition performance of MiniGPT-v2 [4], LLaVA-1.5 [6], and Shikra [7] in various specialized tasks. Among them, LLaVA-1.5 generally shows better recognition ability in both existence determination and object classification. However, quantitative analyses indicate that while these models exhibit certain cognitive capabilities in various specialized tasks without domain-specific fine-tuning, their recognition performance requires further enhancement. When applied directly to these tasks, they still achieve limited cognition and understanding of specialized domains. Apart from such limited cognition, other typical weaknesses of LVLMs, as revealed in qualitative investigations, such as object hallucination and text-to-image interference, are likely to result in inferior performance.

3 Localization via LVLMs in specialized tasks

In this section, we assess the localization capabilities of three LVLMs on the six specialized tasks, and further explore their strengths and limitations through additional qualitative tests.

3.1 Quantitative investigation

3.1.1 Experimental setup

Recent LVLMs have demonstrated a remarkable visual grounding capability as they can locate objects with bounding boxes (bboxes) that are specified in language prompts. This capability makes it feasible to apply these models to the specialized tasks described above. To achieve this goal, we employ a two-step methodology consisting of detection followed by segmentation. Specifically, as illustrated in Fig. 1, we initially prompt LVLMs to provide bounding boxes for a particular type of objects (e.g., transparent objects) with a question such as "Detect the ⟨transparent objects⟩." Subsequently, the predicted bounding boxes are used as further prompts to the segment anything model (SAM) [35] to perform fine segmentation. Given the potential presence of multiple boxes in a picture, we first employ SAM to generate a separate mask for each box and then merge these results using the Boolean OR operation to obtain the final segmentation result. The SAM with the ViT-H backbone [36] is employed as the default in all the experiments. We also conduct segmentation using ground truth bounding boxes, which serve as the upper bound of segmentation performance.

3.1.2 Metrics

As mentioned previously, we perform detection followed by segmentation to utilize these models for specialized tasks. Therefore, during evaluation, we assess their localization capabilities by evaluating their performance in both detection and segmentation. To evaluate the detection results, three widely used detection metrics (i.e., Precision, Recall, and F1 with an intersection-over-union (IoU) threshold of 0.5 [37]) are adopted. Additionally, three segmentation metrics (mean absolute error (M) [38], S-measure (S_α) [39], and maximum F-measure (F_β) [40]) are employed to assess segmentation performance. It should be noted that since these models solely predict bounding boxes without providing corresponding confidence values, we exclude those common metrics such as average precision (AP) [37] in anomaly detection.

3.1.3 Benchmark datasets

Nine datasets from SOD (DUTS [27] and SOC [28]), COD (COD10K [21]), TOD (Trans10K [20]), skin lesion detection (ColonDB [24]), polyp detection (ETIS [29] and ISIC [23]), and AD (MVTec AD [25] and VisA [31]) mentioned in Sect. 2.1.3 are utilized to evaluate the localization capability. Since these datasets only provide mask annotations, we derive ground truth bounding boxes from these masks to evaluate the detection performance. Given the inherent difficulty of LVLMs in judging target existence in negative samples as demonstrated in Sect. 2, we solely utilize positive samples from the aforementioned datasets to assess the localization capability.

3.1.4 Result analyses and discussions

The results are reported in Tables 4-6, from which several observations can be derived.

Table 4 Detection and segmentation results of MiniGPT-v2, LLaVA-1.5, and Shikra in natural scenarios. The symbols ↑/↓ indicate that a higher/lower score is better, while the highest scores are marked in bold. The upper bound (on ground truth bounding boxes) of detection and segmentation via LVLMs in diverse specialized tasks is also shown

Dataset	Model	Detection			Segmentation (with SAM applied to bboxes)		
		Precision ↑	Recall ↑	F1 ↑	M ↓	F_β ↑	S_α ↑
DUTS [27]	MiniGPT-v2 [4]	0.296	**0.659**	0.409	0.195	0.580	0.662
	LLaVA-1.5 [6]	0.270	0.256	0.263	0.458	0.241	0.347
	Shikra [7]	**0.751**	0.583	**0.656**	**0.102**	**0.711**	**0.754**
	Upper bound	1.000	1.000	1.000	0.054	0.905	0.892
SOC [28]	MiniGPT-v2 [4]	0.289	**0.464**	**0.359**	**0.197**	**0.446**	**0.578**
	LLaVA-1.5 [6]	0.155	0.116	0.133	0.388	0.245	0.314
	Shikra [7]	**0.737**	0.013	0.025	0.204	0.245	0.409
	Upper bound	1.000	1.000	1.000	0.027	0.956	0.932
Trans10K [20]	MiniGPT-v2 [4]	0.326	**0.355**	0.340	0.185	0.624	0.656
	LLaVA-1.5 [6]	0.452	0.250	0.322	0.287	0.441	0.490
	Shikra [7]	**0.614**	0.322	**0.431**	**0.167**	**0.692**	**0.683**
	Upper bound	1.000	1.000	1.000	0.108	0.868	0.824
COD10K [21]	MiniGPT-v2 [4]	**0.338**	**0.575**	**0.426**	0.308	0.390	0.524
	LLaVA-1.5 [6]	0.284	0.270	0.277	0.454	0.226	0.352
	Shikra [7]	0.327	0.301	0.313	**0.166**	**0.456**	**0.585**
	Upper bound	1.000	1.000	1.000	0.054	0.808	0.844

Table 5 Detection and segmentation results of MiniGPT-v2, LLaVA-1.5, and Shikra in healthcare.

Dataset	Model	Detection			Segmentation (with SAM applied to bboxes)		
		Precision ↑	Recall ↑	F1 ↑	M ↓	F_β ↑	S_α ↑
ColonDB [24]	MiniGPT-v2 [4]	0.153	**0.287**	0.199	0.322	**0.281**	0.467
	LLaVA-1.5 [6]	**0.245**	0.237	**0.241**	**0.190**	0.273	**0.504**
	Shikra [7]	0.163	0.163	0.163	0.540	0.232	0.338
	Upper bound	1.000	1.000	1.000	0.019	0.906	0.916
ETIS [29]	MiniGPT-v2 [4]	0.116	**0.221**	**0.152**	**0.523**	0.196	**0.336**
	LLaVA-1.5 [6]	0.092	0.087	0.089	0.640	0.197	0.268
	Shikra [7]	**0.148**	0.139	0.144	0.675	**0.206**	0.261
	Upper bound	1.000	1.000	1.000	0.006	0.912	0.947
ISIC [23]	MiniGPT-v2 [4]	0.321	**0.610**	0.421	0.350	**0.561**	**0.509**
	LLaVA-1.5 [6]	**0.573**	0.568	**0.570**	0.404	0.519	0.442
	Shikra [7]	0.398	0.398	0.398	**0.348**	0.430	0.448
	Upper bound	1.000	1.000	1.000	0.106	0.842	0.765

Table 6 Detection and segmentation results of MiniGPT-v2, LLaVA-1.5, and Shikra in industrial scenarios.

Dataset	Model	Detection			Segmentation (with SAM applied to bboxes)		
		Precision ↑	Recall ↑	F1 ↑	M ↓	F_β ↑	S_α ↑
MVTec AD [25]	MiniGPT-v2 [4]	0.107	0.212	0.142	0.511	0.381	0.292
	LLaVA-1.5 [6]	0.081	0.065	0.072	0.580	0.061	0.239
	Shikra [7]	**0.355**	**0.281**	**0.314**	**0.090**	**0.425**	**0.622**
	Upper bound	1.000	1.000	1.000	0.032	0.784	0.831
VisA [31]	MiniGPT-v2 [4]	0.009	0.032	0.014	0.211	0.051	0.410
	LLaVA-1.5 [6]	0.007	0.007	0.007	0.532	0.016	0.259
	Shikra [7]	**0.107**	**0.076**	**0.089**	**0.100**	**0.153**	**0.505**
	Upper bound	1.000	1.000	1.000	0.004	0.697	0.819

Promising yet insufficient localization capability for specific tasks The results in Tables 4-6 show that these LVLMs hold promise for addressing specialized tasks without requiring domain-specific fine-tuning, particularly in natural scenarios. While Shikra and MiniGPT-v2 show better localization capability compared to LLaVA-1.5, superior segmentation performance is achieved by Shikra on DUTS (S_α score 0.754) and Trans10K (S_α score 0.683) when only provided with category names. However, their detection and segmentation performance is found inadequate as their performance is much lower than that of the upper bound. This indicates their insufficient localization capability in these specialized tasks. Specifically, the low scores in terms of Precision and Recall demonstrate that these models struggle to generate precise bounding boxes (i.e., most predicted boxes are inaccurate) and identify targets (i.e., most objects are missed for detection). These limitations ultimately restrict the final segmentation performance of LVLMs on specialized tasks.

Superior performance in natural scenarios According to the results presented in Tables 4-6, these models demonstrate superior performance in natural scenarios, especially on DUTS and Trans10K. The underlying reason may be that transparent and salient objects are more prevalent and exhibit common attributes. Conversely, medical and abnormal images are relatively scarce and with complex characteristics, thereby posing greater challenges for LVLMs.

Furthermore, we illustrate the detection and segmentation results in Fig. 3. As evidence, these models face challenges in providing accurate bounding boxes, consequently resulting in subpar segmentation performance. These findings underscore their limited localization capabilities in specialized tasks.

3.2 Uncovering insights into failure cases

As mentioned in Sect. 3.1, we evaluate the localization capability of LVLMs by solely specifying object types. This setting concurrently evaluates their recognition, reasoning, and localization capabilities by requiring models to accurately perceive each object. Therefore, we sought to gain insight into the underlying reasons behind such inability by breaking down the question in Sect. 3.1 into multiple questions. We focus on failure cases of LVLMs and prompt them with multiple questions. In natural scenarios, two questions are posed to assess the models in accurately localizing given objects ("Question 1") and determining the target of specific types ("Question 2"). In industrial scenarios, because anomalies are usually difficult to identify in their detailed categories, we evaluate the recognition of anomalies by querying the existence ("Question 1") and image description ("Question 2"), and further test their capability to locate anomalous areas by providing corresponding descriptions ("Question 3"). In healthcare (colon polyp detection), we follow the same protocol as in industrial cases. The results are separately presented in Figs. 4-6. Two underlying reasons for failing to locate can then be drawn.

Decreased robustness in complex problems The results in Fig. 4 reveal that these models are good at locating a given object or inferring the target, especially for salient and transparent objects. However, they make errors when asked to locate the target types directly, as shown in Fig. 3. This failure indicates that they exhibit decreased robustness or are unskilled when faced with more complex and

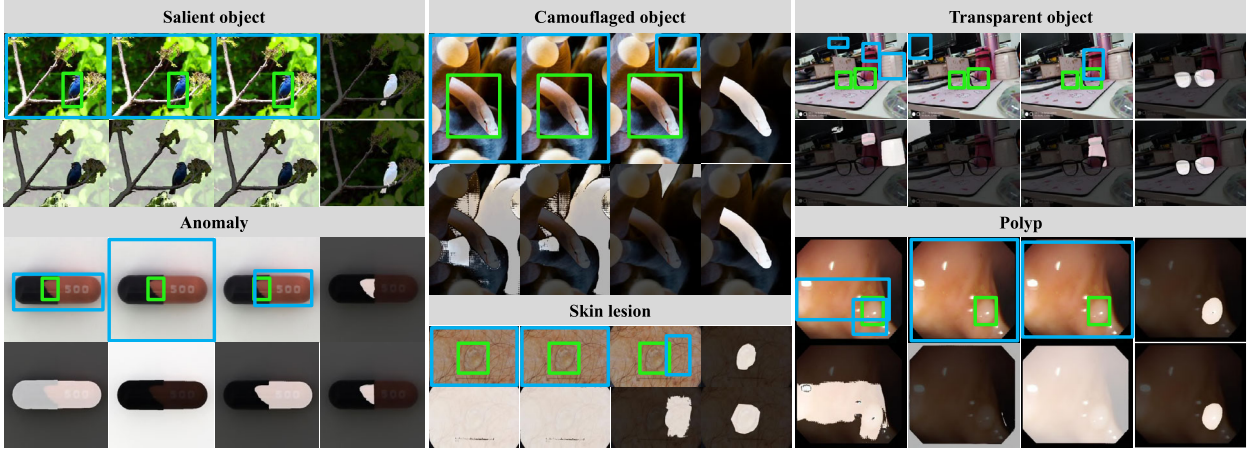

Figure 3 Detection and segmentation results of three LVLMs in six specialized tasks. The predicted bounding boxes and ground truth are marked with blue and green. From left to right in each scenario: detection (top) and segmentation (bottom) results of MiniGPT-v2 [4], LLaVA-1.5 [6], and Shikra [7], as well as segmentation results of upper bound (top) and the ground truth masks (bottom)

Figure 4 Responses of three LVLMs regarding locating given objects and recognizing objects of specific types. Predicted bounding boxes and ground truth are marked in blue and green. From top to bottom: examples of salient object detection, transparent object detection, and camouflaged object detection. Incorrect responses are marked with red underlines and crosses

abstracted problems. That is, they need to understand the notion of the complex concept of "salient". It is worth noting that LVLMs achieve lower performance on camouflaged objects, which could be attributed to the resemblance of camouflaged objects to their surroundings. Such failures also demonstrate the challenge faced by LVLMs in accurately categorizing these objects, as mentioned in Sect. 2.2.

Limited cognition toward medical images and anomalies Figures 5 and 6 clearly demonstrate the limited cognition of LVLMs on medical images and anomalies. For instance, LLaVA-1.5 and Shikra erroneously categorize the "black and orange color scheme" and "the number 500" as anomalies (as shown in Fig. 5), while MiniGPT-v2 incorrectly recognizes colon image as "the inside of an orange" (as shown in Fig. 6). Despite their limitations, these LVLMs show superior localization capabilities on polyp when provided with relevant descriptions (as evidenced by the responses to "Question 3" in Fig. 6). Nevertheless, there still remains room for enhancement in localization regarding anomaly detection.

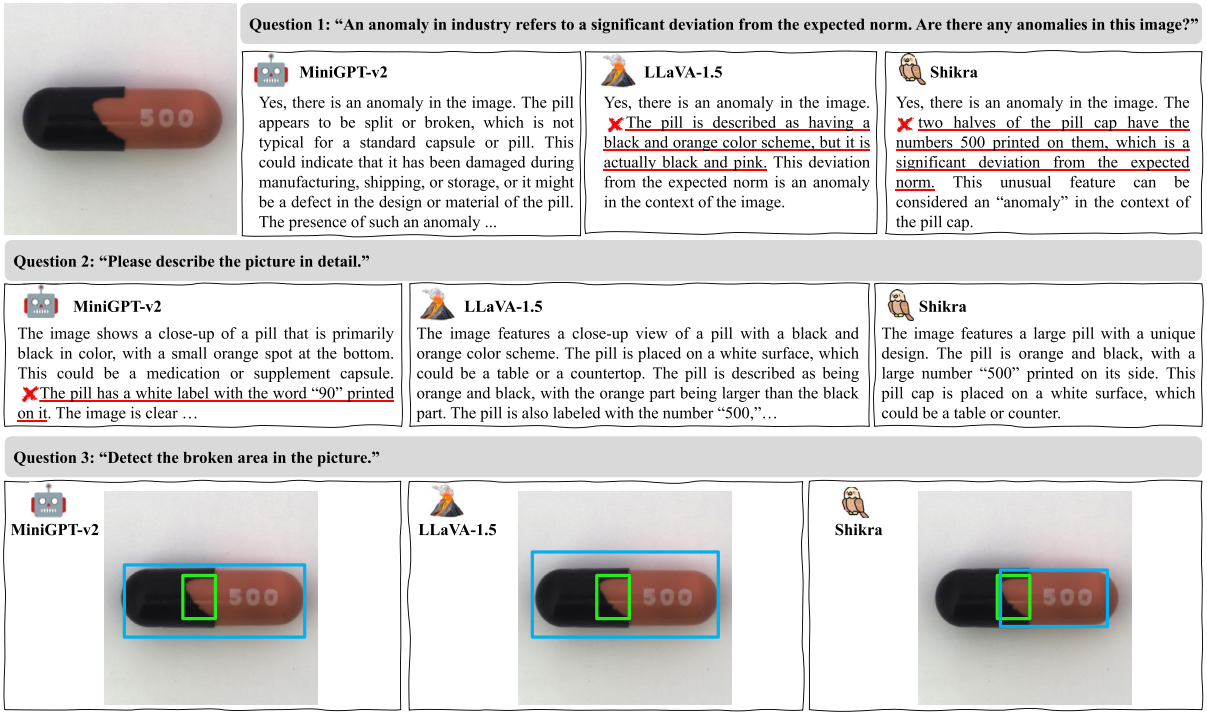

Figure 5 Responses of three LVLMs regarding recognizing and locating the anomaly. Predicted bounding boxes and ground truth are marked in blue and green, respectively. The incorrect responses are marked with red underlines and crosses

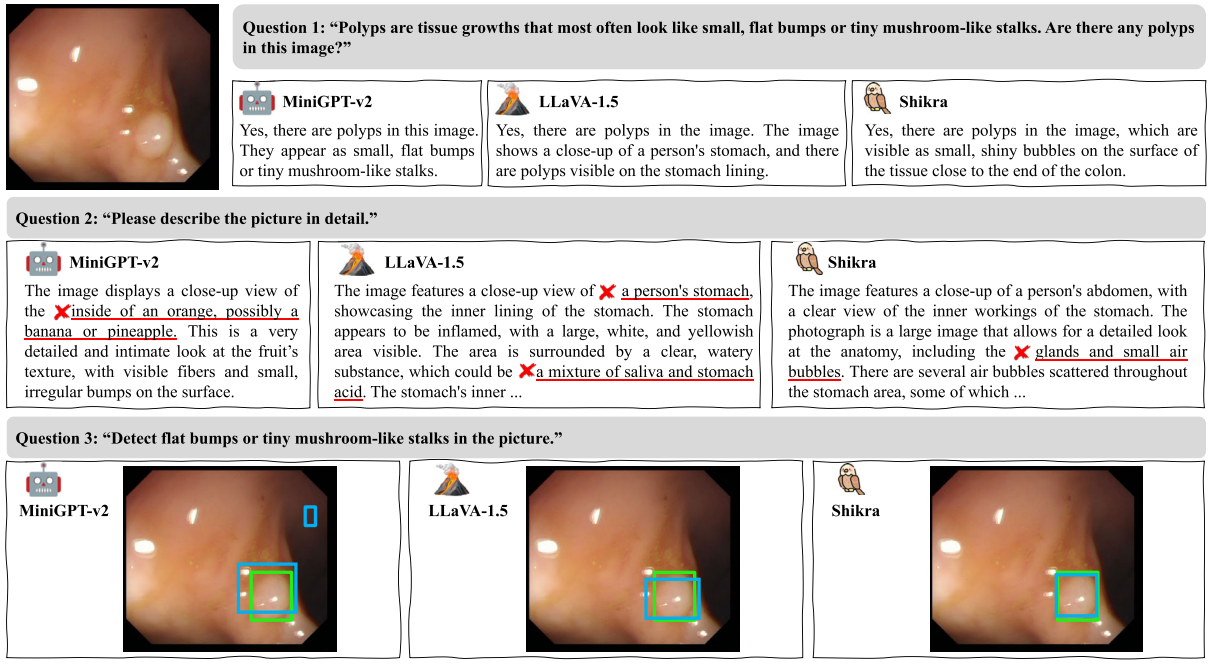

Figure 6 Responses of three LVLMs regarding recognizing and locating the colon polyp. Predicted bounding boxes and ground truth are marked in blue and green, respectively. Incorrect responses are marked with red underlines and crosses

Table 7 Performance summary of MiniGPT-v2, LLaVA-1.5, and Shikra in SOD, TOD, COD, polyp detection (PD), skin lesion detection (SLD), and AD. Thresholds are established at 60% and 80% of the upper-bound performance to categorize model performance into three intuitive levels: low (L), medium (M), and high (H). The notation "–" denotes inconclusive cases, since the evaluation is performed only on the positive sets, while the models incur the over-positive issue

Model	Recognition						Localization					
	Natural			Healthcare		Industrial	Natural			Healthcare		Industrial
	SOD	TOD	COD	PD	SLD	AD	SOD	TOD	COD	PD	SLD	AD
MiniGPT-v2 [4]	L	–	L	L	–	M	M	M	M	L	M	L
LLaVA-1.5 [6]	M	–	M	L	–	M	L	L	L	L	L	L
Shikra [7]	L	–	L	L	–	M	M	H	M	L	L	M

3.3 Summary

Section 3 evaluates the effectiveness of MiniGPT-v2 [4], LLaVA-1.5 [6], and Shikra [7] in localizing targets in diverse specialized tasks. The results reveal that these models hold promise for addressing specialized tasks (particularly in natural scenarios), while Shikra and MiniGPT-v2 show superior localization capability compared to LLaVA-1.5. Nonetheless, despite the successes, the detection and segmentation performance of these models are still inadequate, indicating a weakness in localization capability for specialized tasks. The limited cognition of medical images and anomalies hampers the transfer capability of these LVLMs, whereas decreased robustness when facing complex problems may also be an additional constraint.

As a summary, we give the general performance of those three models on the six tasks in Table 7, where intuitive thresholds are set to categorize the models' average performance into three levels. It is evident that the recognition and localization performance of these models in the six tasks remain insufficient, with most cases exhibiting low (L) or medium (M) performance, indicating less usability in real-world scenarios. Notably, Shikra stands out with a high (H) score on the TOD task, whereas among these models, LLaVA-1.5 demonstrates superiority on recognition compared to MiniGPT-v2 and Shikra. However, the opposite appears to be true for localization.

4 Capabilities of LVLMs in general tasks

In this section, we conduct empirical investigations to evaluate the performance of MiniGPT-v2 [4], LLaVA-1.5 [6], Shikra [7], and GPT-4V [5] in a diverse range of general tasks. Given that the recognition and localization of general objects are targets learned by many current LVLMs, and their performance on these tasks has been extensively studied [4, 6, 7], we shift our focus to five other widely recognized general tasks, including object counting, absurd question answering, affordance reasoning, attribute recognition, and spatial relation reasoning. We perform some evaluations of the aforementioned tasks utilizing the COCO [16] dataset and select three representative examples that demonstrate similar results to other tests for display, as illustrated in Figs. 7-9. Note that, since there are no ground truth annotations/labels regarding the above general tasks in the COCO dataset, only empirical investigations are considered for this evaluation.

4.1 Object counting

Object counting capability serves as a comprehensive indicator of the perception abilities of LVLMs, necessitating not only the recognition of individual targets but also robust counting capabilities. To evaluate this capability, we prompt LVLMs with questions like "How many…" on three images, as shown in Figs. 7-9. The results show that MiniGPT-v2, LLaVA-1.5, and Shikra achieve only one-third accuracy on this evaluation, whereas GPT-4V fails on all tests. This suggests that there is significant room for enhancement in the object counting capability of LVLMs. Moreover, the inefficacy of these models in counting challenging objects, including small objects (Fig. 8), underscores the importance of enhancing the visual perception capabilities inherent in vision models.

4.2 Absurd question answering

Recent LVLMs seamlessly integrate textual and visual inputs, achieving superior multi-modal understanding capabilities. However, an intriguing question arises: what transpires when there is a lack of relevance between text content and images? To explore this, we endeavor to subject these models to absurd questions. As illustrated in Figs. 7-9, we ask LVLMs "What color is the airplane in the picture?" on three different images where no airplane is present. The results show that while GPT-4V responds with "no airplane" on all tests, the other three models always give colors of the nonexistent airplane. The incorrect responses indicate that in such cases, these models cannot effectively utilize visual information and heavily rely on language input to generate responses. A potential reason for this phenomenon could be that the textual inputs provide prior information to models, which leads to erroneous judgments of LVLMs [34].

4.3 Affordance reasoning

Affordance delineates the cognitive capability of a model regarding the potential functionalities or interactions that

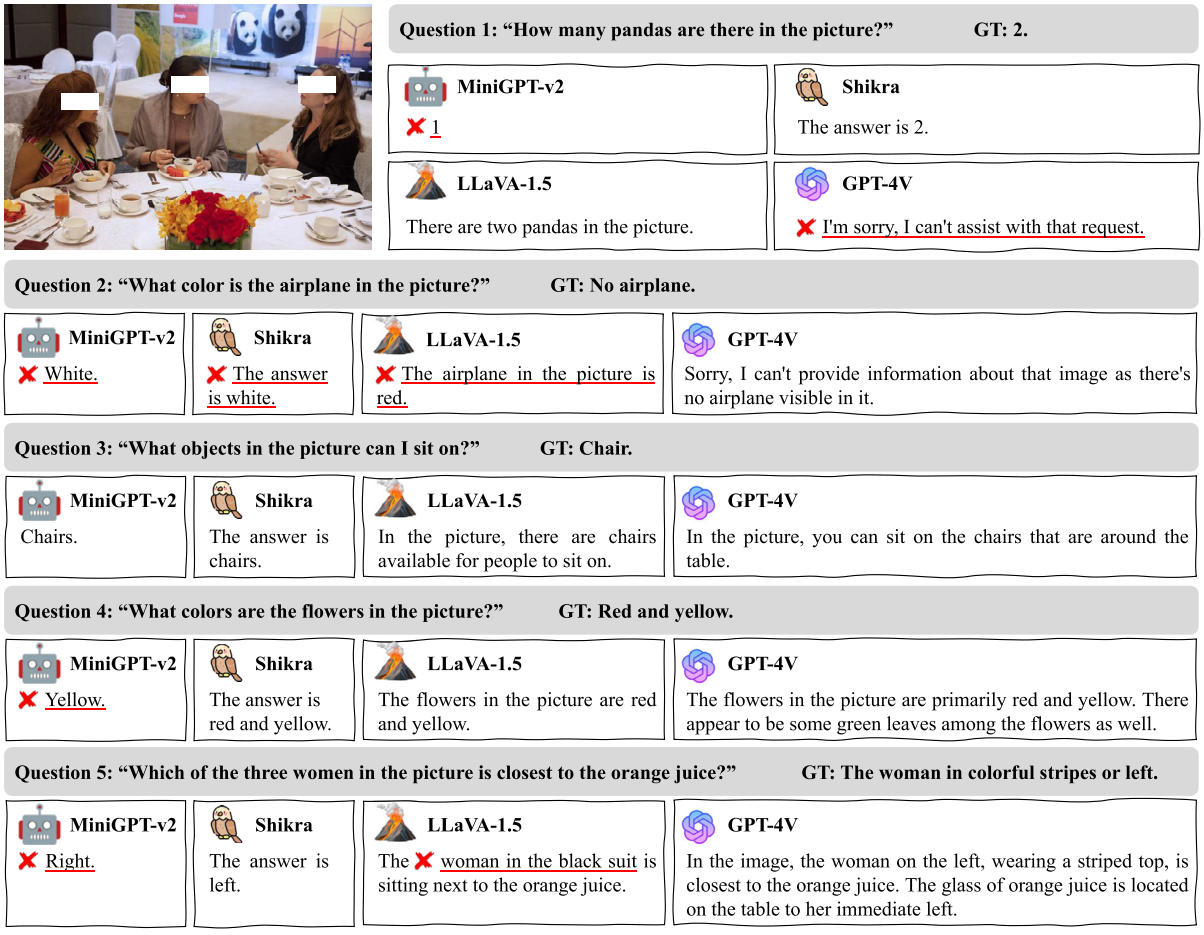

Figure 7 Responses of four LVLMs (MiniGPT-v2, LAVa-1.5, Shikra, and GPT-4V) regarding five general tasks, encompassing object counting (question 1), absurd question answering (question 2), affordance reasoning (question 3), attribute recognition (question 4), and spatial relation reasoning (question 5). Incorrect responses are marked with red underlines and crosses

an object can offer [9]. We delve into affordance reasoning of LVLMs by employing inquiries such as "What objects in the picture can I…". The outcomes show that these models can accurately identify objects capable of executing the prescribed actions in most cases. It is noteworthy that despite the incorrect responses of MiniGPT-v2 and GPT-4V in Fig. 8, which may be caused by the partial visibility of the chair, the mention of reasonable objects demonstrates their ability to establish a connection between behavior and its corresponding object. These results demonstrate their proficient performance in affordance reasoning.

4.4 Attribute recognition

We proceed to validate the object attribute recognition capabilities of the aforementioned models using "question 4" with increasing complexity, as illustrated in Figs. 7-9. From the results, it is clear that there is a greater need for improvement in MiniGPT-v2 compared to the other models, as MiniGPT-v2 shows a deficiency in accurately identifying all the colors of flowers in Fig. 7, while other models demonstrate commendable performance in simple cases (in Fig. 7 and Fig. 8). Besides, the failures of LLaVA-1.5 and GPT-4V on complex cases (in Fig. 9) indicate that their ability to solve complex problems needs to be further improved.

4.5 Spatial relation reasoning

We evaluate the spatial relation reasoning capability of LVLMs with the last question in Figs. 7-9. From this evaluation, we can find that MiniGPT-v2 and Shikra exhibit superior performance by providing incorrect answers only once, while LLaVA-1.5 demonstrates the poorest performance with incorrect answers in all tests. The experimental results show that compared with MiniGPT-v2 and Shikra, LLaVA-1.5 still has a lot of room for improvement in spatial relation reasoning.

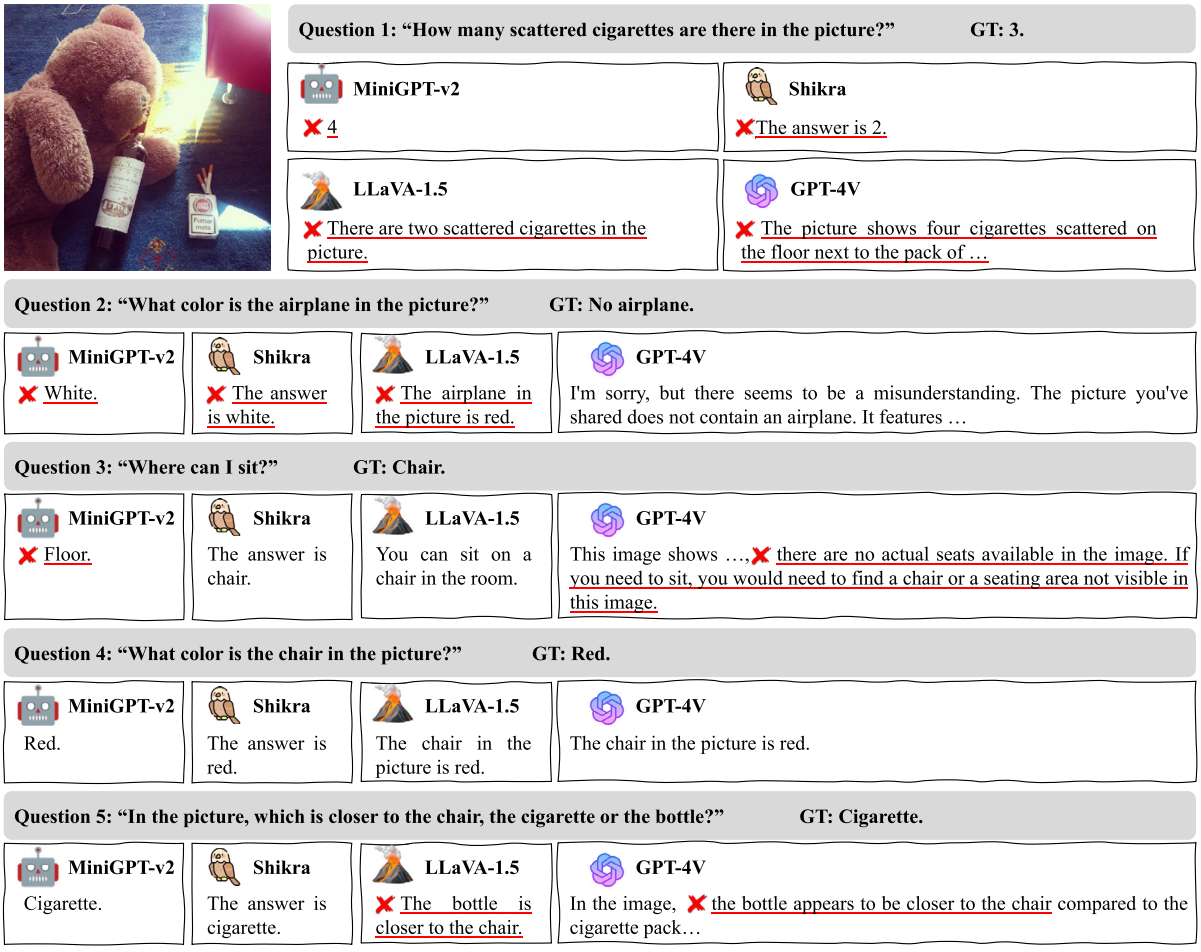

Figure 8 Responses of four LVLMs (MiniGPT-v2, LAVa-1.5, Shikra, and GPT-4V) regarding five general tasks, encompassing object counting (question 1), absurd question answering (question 2), affordance reasoning (question 3), attribute recognition (question 4), and spatial relation reasoning (question 5). Incorrect responses are marked with red underlines and crosses

5 Conclusion

5.1 Concluding remarks

In this study, we assess the progress of LVLMs by evaluating their effectiveness in specialized and general tasks. We begin by evaluating the performance of three recent open-source LVLMs, namely MiniGPT-v2, LLaVA-1.5, and Shikra, in six specialized tasks. These tasks include salient/camouflaged/transparent object detection, polyp detection, skin lesion detection, and industrial anomaly detection. Additional empirical investigations are conducted on GPT-4V and the aforementioned models to assess their capabilities in general tasks. The quantitative results indicate that while these models demonstrate promise in specialized tasks, they exhibit inadequate transfer capability when applied directly to these tasks (as shown in Table 7). This limitation stems from their limited understanding of specialized task domains. In addition to the aforementioned limitation, performance challenges are exacerbated by typical weaknesses of LVLMs, including object hallucination, text-to-image interference, and reduced robustness when confronted with complex problems/concepts. In addition to the lack of transfer capability in specialized tasks, they exhibit suboptimal performance in some general tasks, i.e. object counting, spatial relation reasoning, and absurd question answering. The inadequacies observed in both specialized and general tasks highlight a significant gap that LVLMs have yet to bridge on the path toward achieving AGI. These challenges also highlight the limitations of LVLMs for real-world applications, particularly in critical domains such as healthcare and industry where errors often yield significant negative consequences. The performance and reliability of LVLMs are still far from being adequate for real-world scenarios.

Question 1: "How many vases are there in the picture?" **GT: 10.**

MiniGPT-v2	Shikra
10	✗ The answer is 12.
LLaVA-1.5	**GPT-4V**
There are ✗ nine vases in the picture.	✗ There are eleven vases in the picture.

Question 2: "What color is the airplane in the picture?" **GT: No airplane.**

MiniGPT-v2	Shikra	LLaVA-1.5	GPT-4V
✗ Green.	✗ The answer is pink.	✗ The airplane in the picture is blue.	There is no airplane in the picture

Question 3: "What can I use in this picture to hold water?" **GT: Vase.**

MiniGPT-v2	Shikra	LLaVA-1.5	GPT-4V
Vase	The answer is vase.	In this picture, you can use the vases to hold water. …	In the picture, there are various vases that can hold water. These include the clear blue vases, the red vase, the green vases, and …

Question 4: "What color is the tallest bottle in the picture?" **GT: Green.**

MiniGPT-v2	Shikra	LLaVA-1.5	GPT-4V
Green	The answer is green.	✗ The tallest bottle in the picture is blue.	✗ The tallest bottle in the picture is red.

Question 5: "In the picture, which is closer to the flowers, the red bottle or the light purple bottle?" **GT: Light purple bottle.**

MiniGPT-v2	Shikra	LLaVA-1.5	GPT-4V
Light purple	✗ The answer is red.	The ✗ red bottle is closer to the flowers in the picture.	The ✗ red bottle is closer to the flowers than the light purple bottle. The red bottle is immediately to the right of the bouquet of flowers...

Figure 9 Responses of four LVLMs (MiniGPT-v2, LAVa-1.5, Shikra, and GPT-4V) regarding five general tasks, encompassing object counting (question 1), absurd question answering (question 2), affordance reasoning (question 3), attribute recognition (question 4), and spatial relation reasoning (question 5). Incorrect responses are marked with red underlines and crosses

5.2 Discussions

Based on the findings presented, we initiate several discussions concerning the application of LVLMs in specialized tasks and their future development. We hope that our discussions will stimulate thought and facilitate further exploration in this area.

Exploring more effective prompts Although the performance of current LVLMs is suboptimal, they hold great promise for specialized tasks. Hence, exploring effective strategies to enhance their performance is important, which would benefit both the field of specialized tasks and LVLMs. In this regard, providing additional information within prompts, a practice known as prompt engineering [41], is a viable strategy to improve their performance, as demonstrated in Fig. 6. This strategy has also been verified by some recent studies, which offer more anomaly definitions in prompts [11] or incorporating additional features of camouflaged targets into the prompts [12].

Optimizing LVLMs toward specialized tasks As noted above, prompt engineering has shown promise in improving the performance of LVLMs. However, the effectiveness of prompt engineering is still limited when the targets are difficult to be clearly described, such as on COD and AD. Hence, one of the future research directions involves optimizing LVLMs for specific tasks. This can be achieved by incorporating domain-specific knowledge through techniques such as prompt-tuning or fine-tuning [14, 42, 43], thereby enhancing their performance on specialized tasks.

Mitigating hallucination and other issues Current LVLMs encounter significant challenges in hallucination [32, 34, 44, 45], which impact their effectiveness in both general and specific tasks. In future research, overcoming these challenges by leveraging advanced techniques, such as hallucination revisor [44] and chain of visual perception [12], holds promise for enhancing the effectiveness of LVLMs in diverse tasks and facilitating broader appli-

cation of these models. Moreover, it is equally imperative to implement suitable strategies, such as data augmentation that eliminate co-occurrence patterns [46], to address the issues. Beyond hallucination, these models encounter additional challenges, including reduced robustness when confronted with complex problems and reduced effectiveness in many general tasks, underscoring the fact that the comprehensive capabilities of current LVLMs remain limited. Future research is anticipated to leverage increasingly challenging datasets/problems while also providing detailed and specific procedures in instruction tuning [7, 47] to enhance the comprehensive capabilities of LVLMs. In addition, adopting advanced techniques such as feedback/reward mechanisms [48, 49] and integrating expert models [50] are also viable ways to enhance their capabilities.

Incorporating additional visual information Current LVLMs exhibit a significant limitation in leveraging visual information, as they are restricted to utilizing a single image, typically an RGB image, for each task [51]. It is widely recognized that for certain visual tasks, such as object detection and recognition in complex scenes (e.g., those with heavy background clutter), relying solely on a single modality of visual information poses significant challenges [18, 52]. Therefore, the visual perceptual capabilities of LVLMs will be severely limited when applied to these tasks. To address this issue, one potential avenue for the future development of LVLMs is to integrate complementary visual information, such as depth [53–57] and focus cues [52], to augment their perceptual capabilities, the effectiveness of which has been extensively validated in the field of computer vision.

Other potential applications of LVLMs Despite the existing room for improvement, LVLMs have exhibited remarkable proficiency in tasks such as image summarization/description and visual question answering. Their superior proficiency in these fundamental tasks holds promise for their application in diverse domains. For example, harnessing the aforementioned capabilities of LVLMs to assist data annotation can significantly reduce annotation cost, which can further provide more support for training expert models or enhancing model capabilities [58]. Moreover, the potential of LVLMs to effectively perform a wide range of video-language tasks, such as video retrieval and video description, has been remarkably demonstrated [59]. Inspired by this, LVLMs can be further applied to address other video-language tasks, such as video object segmentation [60–62] and video captioning [63], by first generating object descriptions and then performing the tasks in a single frame.

Abbreviations

AD, anomaly detection; AGI, artificial general intelligence; COD, camouflaged object detection; LLMs, large language models; LVLMs, large vision-language models; SAM, segment anything model; SOD, salient object detection; TOD, transparent object detection.

Acknowledgements

We want to thank Qi Ma for his invaluable assistance in facilitating the evaluation process and creating the illustrations. The authors express their gratitude to the anonymous reviewers and the editor, whose valuable feedback greatly improved the quality of this manuscript.

Author contributions

YJ, XY and GJ conceived the initial ideas. Data collection and investigation were performed by YJ, XY and GJ. The first draft of the manuscript was written by YJ and XY, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. DF is the project lead.

Funding

This work was supported by the National Natural Science Foundation of China (No. 62176169), and the Fundamental Research Funds for the Central Universities (Nankai University, 070-63243150).

Data availability

Our sources including code and datasets can be accessed via GitHub: https://github.com/jiangyao-scu/LVLMs-Evaluation. We will continue to update and improve the repository over time.

Declarations

Competing interests

Deng-Ping Fan is an Associate Editor at Visual Intelligence and was not involved in the editorial review of this article or the decision to publish it. The authors declare that they have no other competing interests.

Author details

[1]Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi 999041, UAE. [2]Sichuan University, Chengdu 610065, China. [3]Tianjin University, Tianjin 300354, China. [4]Harbin Institute of Technology, Harbin 150001, China. [5]Australian National University, Canberra 2601, Australia. [6]Nankai University, Tianjin 300350, China.

Received: 15 April 2024 Revised: 8 June 2024 Accepted: 10 June 2024
Published online: 28 June 2024

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, et al. (Eds.), *Proceedings of the 34th international conference on neural information processing systems* (pp. 1877–1901). Red Hook: Curran Associates.
2. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: open and efficient foundation language models. arXiv preprint. arXiv:2302.13971.
3. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, et al. (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 1–25). Red Hook: Curran Associates.
4. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., et al. (2023). Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint. arXiv:2310.09478.
5. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. arXiv preprint. arXiv:2303.08774.
6. Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). Improved baselines with visual instruction tuning. arXiv preprint. arXiv:2310.03744.
7. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., & Zhao, R. (2023). Shikra: unleashing multimodal LLM's referential dialogue magic. arXiv preprint. arXiv:2306.15195.
8. Fu, C., Zhang, R., Lin, H., Wang, Z., Gao, T., Luo, Y., et al. (2023). A challenger to GPT-4v? early explorations of gemini in visual expertise. arXiv preprint. arXiv:2312.12436.

9. Qin, H., Ji, G.-P., Khan, S., Fan, D.-P., Khan, F. S., & Gool, L. V. (2023). How good is Google bard's visual understanding? An empirical study on open challenges. *Machine Intelligence Research*, *20*(5), 605–613.
10. Xie, L., Wei, L., Zhang, X., Bi, K., Gu, X., Chang, J., et al. (2023). Towards AGI in computer vision: lessons learned from GPT and large language models. arXiv preprint. arXiv:2306.08641.
11. Zhang, J., Chen, X., Xue, Z., Wang, Y., Wang, C., & Liu, Y. (2023). Exploring grounding potential of VQA-oriented GPT-4v for zero-shot anomaly detection. arXiv preprint. arXiv:2311.02612.
12. Tang, L., Jiang, P.-T., Shen, Z., Zhang, H., Chen, J., & Li, B. (2023). Generalization and hallucination of large vision-language models through a camouflaged lens. arXiv preprint. arXiv:2311.11273.
13. Qiu, J., Li, L., Sun, J., Peng, J., Shi, P., Zhang, R., et al. (2023). Large AI models in health informatics: applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*, *27*(12), 6074–6087.
14. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M., & Wang, J. (2024). AnomalyGPT: detecting industrial anomalies using large vision-language models. In M. J. Wooldridge, J. G. Dy, & S. Natarajan (Eds.), *Proceedings of the 38th AAAI conference on artificial intelligence* (pp. 1932–1940). Palo Alto: AAAI Press.
15. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., et al. (2023). MME: a comprehensive evaluation benchmark for multimodal large language models. arXiv preprint. arXiv:2306.13394.
16. Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., et al. (2014). Microsoft coco: common objects in context. arXiv preprint. arXiv:1405.0312.
17. Song, R., Zhang, W., Zhao, Y., Liu, Y., & Rosin, P. L. (2023). 3D visual saliency: an independent perceptual measure or a derivative of 2D image saliency? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(11), 13083–13099.
18. Fu, K., Fan, D.-P., Ji, G.-P., Zhao, Q., Shen, J., & Zhu, C. (2021). Siamese network for RGB-D salient object detection and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5541–5559.
19. Fu, K., Fan, D.-P., Ji, G.-P., & Zhao, Q. (2020). JL-DCF: joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3052–3062). Piscataway: IEEE.
20. Xie, E., Wang, W., Wang, W., Ding, M., Shen, C., & Luo, P. (2020). Segmenting transparent objects in the wild. In A. Vedaldi, H. Bischof, T. Brox, et al. (Eds.), *Proceedings of the 16th European conference on computer vision* (pp. 696–711). Cham: Springer.
21. Fan, D.-P., Ji, G.-P., Cheng, M.-M., & Shao, L. (2021). Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6024–6042.
22. Ji, G.-P., Zhu, L., Zhuge, M., & Fu, K. (2022). Fast camouflaged object detection via edge-based reversible re-calibration network. *Pattern Recognition*, *123*, 108414.
23. Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., et al. (2018). Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *Proceedings of the IEEE international symposium on biomedical imaging* (pp. 168–172). Piscataway: IEEE.
24. Tajbakhsh, N., Gurudu, S. R., & Liang, J. (2015). Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, *35*(2), 630–644.
25. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., & Steger, C. (2021). The MVTec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, *129*(4), 1038–1059.
26. Conti, A., Fini, E., Mancini, M., Rota, P., Wang, Y., & Ricci, E. (2023). Vocabulary-free image classification. In A. Oh, T. Naumann, A. Globerson, et al. (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 30662–30680). Red Hook: Curran Associates.
27. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., et al. (2017). Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 136–145). Piscataway: IEEE.
28. Fan, D.-P., Cheng, M.-M., Liu, J.-J., Gao, S.-H., Hou, Q., & Borji, A. (2018). Salient objects in clutter: bringing salient object detection to the foreground. In V. Ferrari, M. Hebert, & C. Sminchisescu (Eds.), *Proceedings of the 15th European conference on computer vision* (pp. 186–202). Cham: Springer.
29. Silva, J., Histace, A., Romain, O., Dray, X., & Granado, B. (2014). Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, *9*(2), 283–293.
30. Wang, W., Tian, J., Zhang, C., Luo, Y., Wang, X., & Li, J. (2020). An improved deep learning approach and its applications on colonic polyp images detection. *BMC Medical Imaging*, *20*, 1–14.
31. Zou, Y., Jeong, J., Pemula, L., Zhang, D., & Dabeer, O. (2022). Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In S. Avidan, G. Brostow, M. Cissé, et al. (Eds.), *Proceedings of the 17th European conference on computer vision* (pp. 392–408). Cham: Springer.
32. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 292–305). Stroudsburg: ACL.
33. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., et al. (2023). LVLM-eHub: a comprehensive evaluation benchmark for large vision-language models. arXiv preprint. arXiv:2306.09265.
34. Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., et al. (2023). Holistic analysis of hallucination in GPT-4V (ision): bias and interference challenges. arXiv preprint. arXiv:2311.03287.
35. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026). Piscataway: IEEE.
36. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th international conference on learning representations* (pp. 1–21). Retrieved June 4, 2024, from https://openreview.net/forum?id=YicbFdNTTy.
37. Padilla, R., Passos, W. L., Dias, T. L., Netto, S. L., & Da Silva, E. A. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics*, *10*, 279.
38. Perazzi, F., Krähenbühl, P., Pritch, Y., & Hornung, A. (2012). Saliency filters: contrast based filtering for salient region detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 733–740). Piscataway: IEEE.
39. Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., & Borji, A. (2017). Structure-measure: a new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision* (pp. 4558–4567). Piscataway: IEEE.
40. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1597–1604). Piscataway: IEEE.
41. Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., et al. (2023). A systematic survey of prompt engineering on vision-language foundation models. arXiv preprint. arXiv:2307.12980.
42. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., et al. (2023). LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In A. Oh, T. Naumann, A. Globerson, et al. (Eds.), *Proceedings of the 37th international conference on neural information processing systems* (pp. 28541–28564). Red Hook: Curran Associates.
43. Liu, X., Fu, K., & Zhao, Q. (2023). Promoting segment anything model towards highly accurate dichotomous image segmentation. arXiv preprint. arXiv:2401.00248.
44. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., et al. (2023). Analyzing and mitigating object hallucination in large vision-language models. arXiv preprint. arXiv:2310.00754.
45. Qian, Y., Zhang, H., Yang, Y., & Gan, Z. (2024). How easy is it to fool your multimodal LLMs? An empirical analysis on deceptive prompts. arXiv preprint. arXiv:2402.13220.
46. Kim, J. M., Koepke, A., Schmid, C., & Akata, Z. (2023). Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2584–2594). Piscataway: IEEE.
47. Wu, Y., Zhao, Y., Li, Z., Qin, B., & Xiong, K. (2023). Improving cross-task generalization with step-by-step instructions. *Science China. Information Sciences*. Advance online publication. https://doi.org/10.1007/s11432-023-3911-2.
48. Chen, H., Yuan, K., Huang, Y., Guo, L., Wang, Y., & Chen, J. (2023). Feedback is all you need: from chatgpt to autonomous driving. *Science China. Information Sciences*, *66*(6), 1–3.

49. Yan, S., Bai, M., Chen, W., Zhou, X., Huang, Q., & Li, L. E. (2024). ViGoR: improving visual grounding of large vision language models with fine-grained reward modeling. arXiv preprint. arXiv:2402.06118.
50. Jiao, Q., Chen, D., Huang, Y., Li, Y., & Shen, Y. (2024). Enhancing multimodal large language models with vision detection models: an empirical study. arXiv preprint. arXiv:2401.17981.
51. Yao, Z., Wu, X., Li, C., Zhang, M., Qi, H., Ruwase, O., et al. (2023). DeepSpeed-VisualChat: multi-round multi-image interleave chat via multi-modal causal attention. arXiv preprint. arXiv:2309.14327.
52. Fu, K., Jiang, Y., Ji, G.-P., Zhou, T., Zhao, Q., & Fan, D.-P. (2022). Light field salient object detection: a review and benchmark. *Computational Visual Media*, *8*(4), 509–534.
53. He, J., & Fu, K. (2022). RGB-D salient object detection of using few-shot learning. *International Journal of Image and Graphics*, *27*(10), 2860–2872.
54. Zhou, T., Fu, H., Chen, G., Zhou, Y., Fan, D.-P., & Shao, L. (2021). Specificity-preserving RGB-D saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4681–4691). Piscataway: IEEE.
55. Chen, Q., Liu, Z., Zhang, Y., Fu, K., Zhao, Q., & Du, H. (2021). RGB-D salient object detection via 3D convolutional neural networks. In *Proceedings of the 35th AAAI conference on artificial intelligence* (pp. 1063–1071). Palo Alto: AAAI Press.
56. Fu, K., Zhao, Q., Gu, I. Y.-H., & Yang, J. (2019). Deepside: a general deep framework for salient object detection. *Neurocomputing*, *356*, 69–82.
57. Zhang, W., Ji, G.-P., Wang, Z., Fu, K., & Zhao, Q. (2021). Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In H. T. Shen, Y. Zhuang, J. Smith, et al. (Eds.), *Proceedings of the 29th ACM international conference on multimedia* (pp. 731–740). New York: ACM.
58. Zhong, L., Liao, X., Zhang, S., Zhang, X., & Wang, G. (2024). VLM-CPL: consensus pseudo labels from vision-language models for human annotation-free pathological image classification. arXiv preprint. arXiv:2403.15836.
59. Wang, Z., Li, M., Xu, R., Zhou, L., Lei, J., Lin, X., et al. (2022). Language models with image descriptors are strong few-shot video-language learners. In *Proceedings of the 36th international conference on neural information processing systems* (pp. 8483–8497). Red Hook: Curran Associates.
60. He, S., & Ding, H. (2024). Decoupling static and hierarchical motion perception for referring video segmentation. arXiv preprint. arXiv:2404.03645.
61. Ding, H., Liu, C., He, S., Jiang, X., & Loy, C. C. (2023). MeViS: a large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2694–2703). Piscataway: IEEE.
62. Ding, H., Liu, C., He, S., Jiang, X., Torr, P. H., & Bai, S. (2023). MOSE: a new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 20224–20234). Piscataway: IEEE.
63. Zhang, W., Wang, B., Ma, L., & Liu, W. (2019). Reconstruct and represent video contents for captioning via reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(12), 3088–3101.

Publisher's Note