Disentangling Pictorial Cue Understanding from Language Bias in VLMs via Depth Ordering Task

Yiqian Liu[1], Iuliia Kotseruba[2], and John K. Tsotsos[1]

[1] York University, Toronto, Canada {yql,tsotsos}@yorku.ca
[2] University of Guelph, Guelph, Canada
kotseruba@uoguelph.ca

Abstract. In this paper, we study depth perception of vision-language models (VLMs) to isolate the effects of pictorial depth cues and disentangle vision and language influences on model performance. To this end, we combine depth-ordering and odd-one-out psychophysical tasks: the VLMs are presented with images where one object is at different depth relative to other, otherwise identical, objects, and must determine whether the odd-one-out target is closer or farther to the observer. To create stimuli, we generate 2D views from simulated and real 3D scenes while controlling the presence of individual pictorial depth cues, enabling a fine-grained analysis of cue-level contributions. Language effects are examined by varying referring expression clarity. We also introduce a novel metric to quantify vision-vs-language sensitivities. Applying this methodology, we create the Odd-One-Out Depth (O3-D) dataset with 37K real and synthetic images and 147K image-question pairs. Evaluation of 12 open-source and commercial models on O3-D shows under-utilization of depth cues and depth-ordering accuracies between 47% and 56%, with no model above chance level. At the same time, our metric reveals strong linguistic bias in the answers. Neither chain-of-thought (CoT) nor in-context learning (ICL) significantly improves performance, suggesting that static image data alone may be insufficient for depth understanding. All code, the image generation pipeline, and the O3-D dataset are publicly released at https://github.com/lyiqian/o3-d.

Keywords: Pictorial Depth Cues · Odd-One-Out · Depth Ordering · Referring Expression Comprehension · VLM · VQA

1 Introduction

Extracting rich spatial structure from images is one of the fundamental computer vision problems. Historically, it has been approached from multiple angles, such as monocular depth estimation [2], object detection [57], segmentation [37], and 3D scene reconstruction [44]. The most recent generation of vision-language models (VLMs) aims to perform scene understanding as a single system.

Despite being trained only on static images, VLMs demonstrate a range of depth-related abilities. For example, past benchmarks [3, 7, 12, 13, 19, 26, 32, 50] showed evidence of VLMs understanding size, distance, and structure of objects.

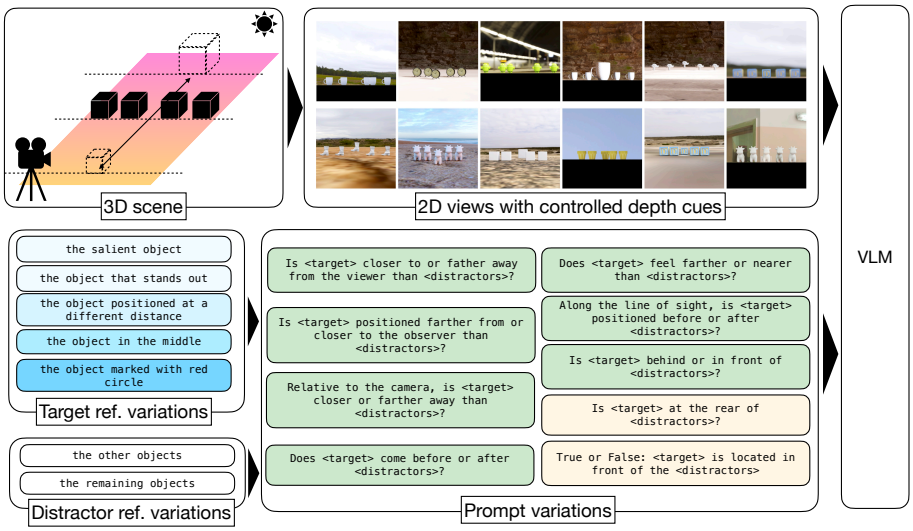


Fig. 1: O3-D probes VLM depth and language understanding. We start by constructing synthetic and real `3D scenes` with diverse backgrounds and objects. Each scene contains 5 objects of the same class, one of which (the target) is of different size and placed at a different depth plane. We then generate a number of `2D views with one or two depth cues` by controlling the camera, light position, *etc*. For each image, we create `prompt variations`: multiple-choice (shown in green) and yes-no (yellow). Within each prompt, we vary the target referring clarity (more specific references are shown in darker shades of blue in the `Target ref. variations` box) and distractor descriptions (`Distractor ref. variations`).

However, utilization of individual depth cues by these models remains understudied. Another challenge in evaluating VLMs is posed by their inherent bimodal nature. As most VLM evaluation protocols are based on Visual Question Answering (VQA), both vision and language modalities interact, making it difficult to disentangle their individual effects on the overall performance.

To address the limitations of existing depth benchmarks, we propose the *Odd-One-Out Depth (O3-D)* dataset and evaluation metrics for systematic evaluation of VLMs' depth understanding capabilities while taking into account complexities arising from the language dimension (Fig. 1). We approach the problem from a psychophysical standpoint by using a depth ordering task [6] within an odd-one-out setup [46]. Specifically, we construct a series of synthetic 3D odd-one-out scenes where one object is at different depth relative to other, otherwise identical, objects (Fig. 2a). We then generate a number of 2D views from these 3D scenes while controlling for 9 common pictorial depth cues [23, 43, 52] to isolate cue-level contributions (Fig. 2c). Additionally, O3-D includes real-world odd-one-out-depth images from two sources: captured in a custom setup and selected from [25]. Along the language dimension, we consider referring expression comprehension, which commonly deals with language ambiguity [18]. Referring

to objects is especially challenging in visual contexts with multiple similar objects, thus we experiment with the referring clarity. In addition, we introduce a novel metric for measuring the relative sensitivities of vision and language input modalities. Lastly, we test whether common techniques such as chain-of-thought (CoT) and in-context learning (ICL) can improve VLMs' depth perception. Our main contributions are summarized as follows:

- We propose a challenging dataset, O3-D for the odd-one-out depth ordering VQA task to evaluate VLMs depth capabilities along vision and language dimensions. For the former, we test utilization of individual pictorial depth cues. For the latter, we vary the clarity of referring expression.
- We design a novel pipeline for constructing synthetic odd-one-out scenes with configurable objects, environments, cues, and prompt variations.
- We formulate a novel metric for measuring the relative influences of vision and language inputs on the VLMs' overall performance.
- Through extensive experimental validation, we demonstrate low individual pictorial cue utilization as well as consistently high language influence across 12 commercial and open-source SOTA VLMs.

2 Related Works

Depth understanding of foundation vision and vision-language models. Many works examined monocular depth understanding of vision models [15, 17, 29, 35, 56] and VLMs [3, 7, 12, 13, 19, 26, 32, 50]. Generally, the depth understanding ability was probed indirectly by testing models' perception of size [12, 13, 15], distance [7, 12, 13], structure [17, 56], and spatial relations [15, 26, 32, 50, 56]. More explicitly, depth perception was tested by depth ordering of points [15, 19, 29], regions [12, 56], or objects [3, 7, 13]. Notably, the effects of the individual depth cues were examined only in DepthCues [15]. The authors gathered images from various datasets, labeled them with 6 pictorial depth cues and tested a range of large vision models to show cue utilization. However, since mostly real-world images were used, cues were not isolated. As a result, 60–80% of images contain occlusion, height-in-plane, relative size and linear perspective cues, whereas other 5 cues appear in fewer than 10% of the images[3]. In contrast, O3-D covers a broader set of 9 depth cues via novel synthetic and real images. This offers a more precise control of the pictorial depth cues as well as excludes the possibility of data leakage (Tab. 1).

Visual question answering (VQA). A recent survey [24] on VQA stressed the importance of properly using visual information against language bias [38] via, for example, weakening language priors or strengthening visual information [20]. Studies [9, 16] also showed that language had a larger impact on the response of VQA. While common 3D VQA datasets [4, 34] contain certain level of question variations, efforts were made to remove unclear and ambiguous questions. O3-D incorporates question variations along many dimensions, including referring

[3] Based on the manual labeling of a random sample of 100 images from [15].

Table 1: Comparison with related datasets. Most existing datasets do not support cue-level analysis, except [15]. O3-D is the only dataset isolating specific cues and cue combinations, allowing direct evaluation of cue utilization.

	Pictorial depth cues				
Dataset	# cues	# present	Controlled	# images	# questions
GeoMeter [3]	n/a	multiple	✗	2,000	11,200
3D-PC [29]	n/a	multiple	✗	7,574	n/a
DepthCues [15]	6	multiple	✗	35,753	n/a
O3-D (ours)	**9**	**1 or 2**	✓	**37,110**	**147,552**

clarity (via general *vs*. specific expressions); together with depth cue variations in O3-D, this allows testing and quantifying vision-language influences on the VQA responses.

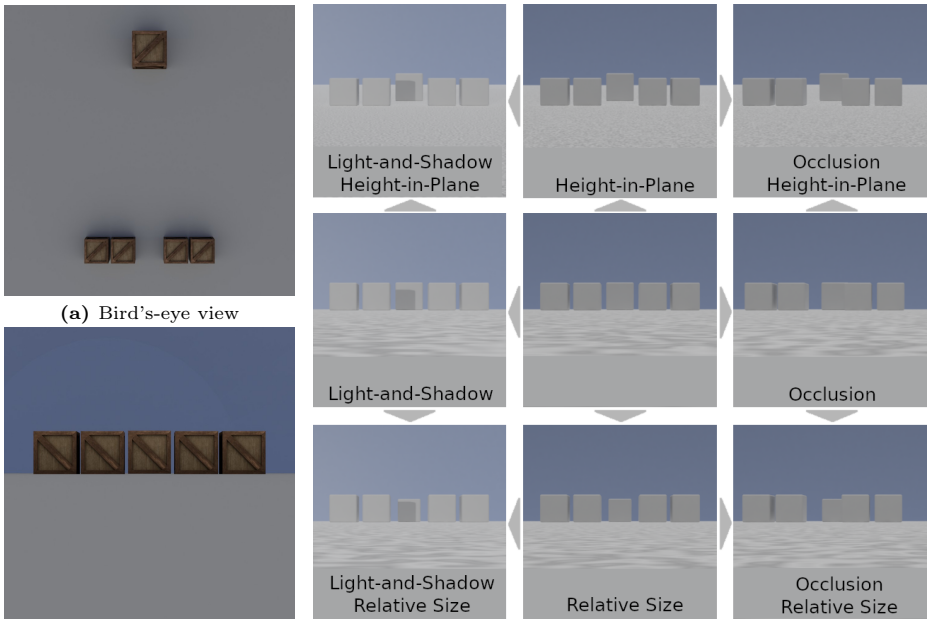
Visual grounding. The essence of visual grounding is associating language description with corresponding visual stimuli [8, 36, 39, 55]. This task is more difficult when multiple objects of a same class are present [30] as wrong associations might impact downstream tasks. Referring expression comprehension (REC) and phrase grounding are the two main tasks of visual grounding, where the former is concerned with fuller descriptions and the latter focuses on shorter phrases [55]. A recent work [14] also explored referring by bounding box coordinates in natural language. The visual grounding capability is helpful for VQA because it encourages the model to consider visual information [24, 39]. In the proposed O3-D, every image contains multiple same-class objects with *similar appearance*, making visual grounding even more challenging.

3 Odd-One-Out Depth Dataset

To systematically evaluate depth perception of VLMs, the proposed O3-D dataset incorporates two well-established psychophysical tasks: odd-one-out [46] and depth ordering [6]. Each image in the O3-D contains multiple similar objects, with only one object (the target) located at a different depth plane relative to other objects (the distractors). To formulate it as a Visual Question Answering (VQA) task, we design a set of prompts for each image with varying referring clarity. Referring expression comprehension [30] allows testing the language abilities of the models. A detailed description of our methodology follows with additional information available in the Supplementary Material.

Pictorial depth cues. To analyze depth perception, we use the following 9 common pictorial cues identified in the psychology literature [23, 43, 52]: Height-in-Plane (HP), Occlusion (OC), Relative Size (RS), Familiar Size (FS), Light-and-Shadow (LS), Texture Gradient (TG), Aerial Perspective/Saturation (SA), Focusness (FO), and Linear Perspective (LP).

Scene construction & cue control. Each O3-D scene contains 1 target and 4 distractors placed on a level surface. The target differs from the distractors only



(a) Bird's-eye view

(b) Ambiguous (base) view

(c) Views with depth cues derived from the base view

Fig. 2: Each 3D scene in O3-D contains 1 target and 4 distractors, where the target is larger (or smaller) and located on a different depth plane (a). When a camera is placed at a certain position, (b) the target appears at the same depth as the distractors. (c) Gray arrows between the images indicate how disambiguated 2D views are generated from the base view (in the center) by adding one or two pictorial depth cues.

in size and is placed at a different depth plane from the distractors (Fig. 2a). Positioning the camera in a certain way (Fig. 2b) creates views with scale ambiguity [48, p. 53]. It is emphasized that scene and view are 3D and 2D concepts, respectively. In other words, from a single 3D scene we generate multiple 2D views with different depth cues. See Fig 1. in Supplementary Material for examples of each.

Fig. 2c illustrates the cue control method and shows some resulting views. From an ambiguous base view (Fig. 2c, center), individual depth cues are added by manipulating the camera, objects, or environment. Translating the camera horizontally and upward introduces the OC and HP cues, respectively; moving the target along the optical axis results in the RS cue; adding directional light so that near objects cast shadows on far ones gives the LS cue; adding textures to objects creates the TG cue; removing textures from the ground controls the LP cue [43]; using a larger camera aperture strengthens the FO cue; common objects with similar shapes but different sizes are used for the FS cue; finally, the natural source of the SA cue is haze, which we simulate by a haze equation [22].

Table 2: Image and question counts in O3-D dataset, grouped by number of pictorial cues present in the images.

Cues	Real imgs	Sim imgs	Img-ques pairs	Reason to include
0-cue	16	1,443	5,796	as negative baseline
1-cue	99	13,746	55,170	controlled single-cue images
2-cue	79	21,556	86,244	controlled two-cue images
mixed	171	-	342	as natural baseline
total	365	36,745	147,552	

While Fig. 2c only shows 4 single-cue and 4 two-cue views, O3-D dataset covers 9 individual cues as well as all possible second-order interactions among them[4].
Synthetic scenes. We use the Kubric [21] simulation environment to render a set of images with the following characteristics:

- the target is randomly scaled to be 10% to 100% larger (smaller), and placed on a farther (nearer) depth plane relative to distractors;
- 37 object classes selected from Kubric assets, with different shape complexities, ranging from simple boxes to complex toys;
- 13 environments with diverse indoor and outdoor backgrounds and realistic ambient lighting;
- 9 individual pictorial cues and 28 pairs of cues. For each cue except FS and LP, we additionally generate images with various cue strengths. For example, cue strengths of HP and OC can be measured by height difference and area of occlusion, respectively.

Overall, we render 13,746 images with single cue and 21,556 with two-cue interactions (see Tab. 2). Also obtained from the Kubric renderer are depth maps, segmentation maps, as well as ground truth target and distractors. More information on the Kubric setup is in Section 2 of Supplementary Material.
Real-world scenes. We set up real-world O3-D scenes in an indoor environment. Following the same procedure as in simulation, multiple views are derived and captured for 3 object classes and 8 cues (excluding Saturation). The 3 object classes are cube, clip, and cup, with the targets being 100%, 25%, and 27% larger, respectively. We use a dark-colored desk as the level surface and a green wall as the background. The above forms our 1-cue and 2-cue real images, as summarized in Tab. 2.
Sensor settings & post-processing. In both simulation and real world, the camera focal length is set close to 50 mm. When the FO cue is unwanted, we turn off depth of field rendering in Blender, or use a small aperture (*e.g.* f/18). We try to reduce undesired variations between shots: for each scene, auto-exposure is run once and turned off; focus is manually set on the center object. The rendered image size is 1024×1024, while real images were resized to 1024×683.

[4] Except Linear Perspective (LP) which has to be tested with Height-in-Plane (HP) because both cues are the result of moving camera above ground.

Table 3: Referring expression variations. As referring to the target object in an O3-D image is challenging, we explore referring expressions with different clarity.

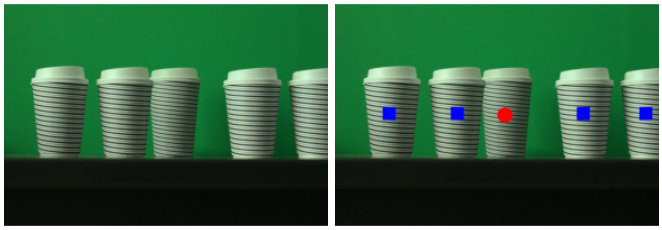
Feature	Clarity	Example expression
saliency	low	Is the *salient object* ...
depth (z)	med	Is the *salient object due to its unique distance* in the image ...
spatial (x, y)	high	Is the *object in the middle* ...
marker	highest	Is the *object marked with a red circle* ...

Image baselines. O3-D contains two special subsets of images: 0-cue and mixed-cue (Tab. 2). The 0-cue set consists of all the ambiguous (base) view images (*e.g.* Fig. 2b) with no pictorial cues. For the mixed-cue set, we select 171 real-world images from O^3 dataset [25] where the odd target was behind or in front of all the distractors. In these images, multiple pictorial cues exist and are not controlled. We label them with two most prominent cues for reference. O^3 images were resized to a max of 1024 pixels in the larger dimension.

Prompt templates. We generate a template-based question space of 1,026 unique prompts, with variations in 4 dimensions: query template, target referring, distractor referring, and response instruction.

Specifically, we obtain 9 templates by collecting depth questions from common VQA datasets [7, 12, 13, 29, 34], and rephrasing them with an LLM [45]. The query templates address various perspectives of depth ordering questions, such as different vocabularies, regular *vs*. yes-no queries, and levels of formalism (see Section 4 in Supplementary Material). Each template also contains two placeholders for target and distractor referring expressions.

As shown in Tab. 3, the target referring expressions vary by their clarity. We explore 4 levels of clarity with distinct referring features. The referring expressions of low and medium clarity require understanding the target as an odd object, described by *e.g.* "salient" or "standing out". Because REC in a context with multiple similar objects is challenging [12, 30], we additionally test a more direct referring mechanism with visual markers placed on objects (Fig. 3b). Our



(a) Resized **(b)** Marked

Fig. 3: Image post-processing. (a) Resize a real image to 1024 × 683. (b) Optionally add markers for easier referring (Tab. 3 bottom row), to gauge the effects of referring expression comprehension (REC) on depth ordering responses.

referring expressions across clarity levels cover all 3 common REC variations of subject, location, and relation [41].

To simplify VLM response parsing, we use a multiple-choice question format, as in [3, 19, 29]. Since only one object (*i.e.* the target) is at a different depth, the depth order can be described with binary answers (*e.g.* closer or farther). Therefore, the response instruction is simply a 2-item option list (*e.g. A. farther. B. closer.*) followed by an instruction (*i.e. Answer A or B.*). The option list is randomized, resulting in normal and reversed orders. We use the forced binary choice because preliminary results showed that if a 'not sure' option was included, VLMs almost always chose it, thus making the analysis less meaningful.

Prompts. For each image in O3-D, we sample *depth ordering* questions with different target referring clarity from the prompt templates. This results in 147K image-question pairs (Tab. 2).

4 Experiment Setup

We run 12 VLMs on O3-D and evaluate their VQA performance on various question formats and pictorial cues.

Baseline. DepthAnythingV2 [54] is selected as the baseline for its robustness to diverse scenes and ability to process high-resolution images. Since DepthAnythingV2 produces depth maps only, we further process the output to make comparisons with VLMs. Specifically, we compute the median depth of target and distractors using their ground truth masks, then determine the depth ordering.

Evaluated VLMs. We evaluate the following VLMs: Kosmos2 [39], PaliGemma2 [47], Qwen2-VL [51], InternVL2.5 [11], DeepSeek-VL [33], LLaVA1.5 [31], VILA1.5 [28], BLIP2 [27], Phi3 [1], Cambrian [49], GPT4.1-mini, and Gemini 2.5 Flash-Lite. All evaluated VLMs are open-source, except GPT and Gemini. Out of these VLMs, only the Cambrian model particularly focuses on utilizing visual information. In addition to language referring, Kosmos2 supports region referring by bounding boxes; results from both referring types will be reported.

Depth-focused SpatialRGPT [12] is not selected because it takes mask referring as part of the input and runs DepthAnything in the background, which is equivalent to our baseline described above.

Metrics. We measure accuracy for all responses. In addition, we introduce the standard deviation of within-group means (SDGM) metric (see Sec. 5.2) to measure VLMs' sensitivities to cue and language variations in the depth ordering task.

In-context learning (ICL) and chain-of-thought (CoT) prompting. For 5 of 12 VLMs, we provide additional few-shot ICL & CoT prompting using the 1-cue, 2-cue, and mixed-cue images (Tab. 2). As image similarity and order matters [5], we retrieve two (target-far and target-near) demonstrations with the same cues as in the main image, in randomized order. Within each demonstration, the image is positioned before the depth question prompt followed by the expected answer [42]. The CoT prompting in the demonstrations only addresses the depth

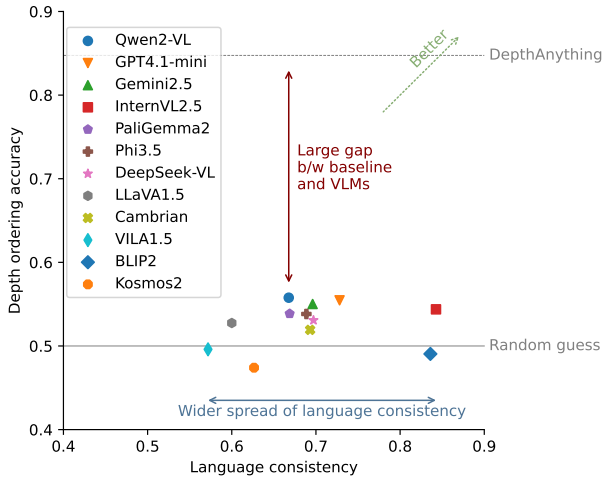


Fig. 4: Performance summary of VLMs on depth ordering VQA. Both depth understanding (y-axis, see Sec. 5.1) and language consistency (x-axis, see Sec. 5.2) can be probed using our O3-D dataset. Depth ordering accuracies of VLMs are close to random guess and inferior to DepthAnythingV2 [54] baseline. VLMs' language consistency has a wide spread.

understanding, not the referring comprehension. As an example, a demonstration can be formatted as follows:

> <image> Is the unique object positioned farther from or closer to the observer than the remaining objects? A. Farther. B. Closer. Answer A or B. (Let's think step by step. The object of interest appears lower than the others. Based on the height-in-plane pictorial cues, it is likely that the object is closer than the other objects.) B.

Response parsing. We parse only the first sentence[5] in a VLM response. Most responses are simply *A* or *B*, and we report accuracy against the ground truth. Responses that do not follow the formatting instruction are scored as positive only if their first sentence contains the correct answer option.

5 Experiment Results

As summarized in Fig. 4, we report the results of experiments that probed VLMs' understanding of pictorial depth cues (Sec. 5.1) and question comprehension (Sec. 5.2). In addition, we discuss cue-level findings and a bias along the near-far spectrum for the vision dimension, and address the consistency of VQA responses for the language dimension. Additional experiment results are presented in Section 6 of Supplementary Material.

[5] CoT prompting occasionally caused extra outputs, which we removed before parsing.

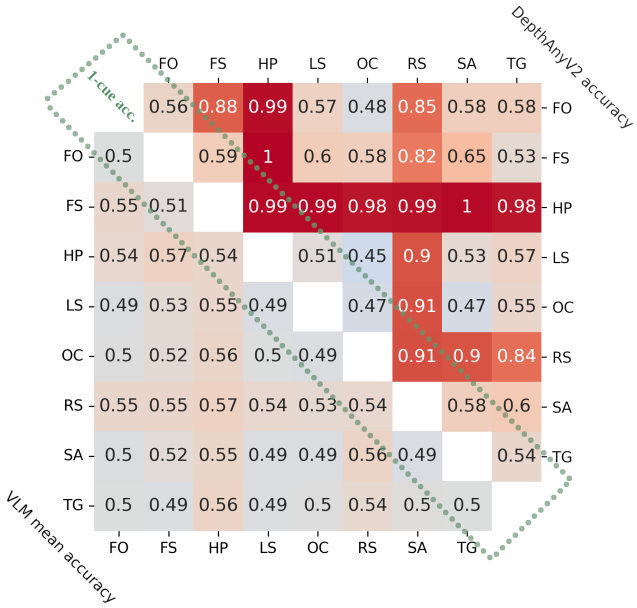


Fig. 5: A combined heatmap of 1-cue and 2-cue mean accuracies of all tested VLMs (bottom-left), *vs*. baseline (top-right). Red- and blue-tinted cells indicate performance above and below chance level (0.5). The two main diagonal cells (within green dotted rectangle) show accuracies for 1-cue depth ordering, whereas the other cells report 2-cue interactions. The depth ordering performance is better whenever HP or RS cue are present. (FO: Focusness, FS: Familiar Size, HP: Height-in-Plane, LS: Light-and-Shadow, OC: Occlusion, RS: Relative Size, SA: Saturation, TG: Texture Gradient)

5.1 Vision Dimension

In order to reduce the interference of referring expression comprehension, the results of vision dimension are reported only for the images with markers, *i.e*. the highest referring clarity in Tab. 3, unless otherwise noted. When comparing with 2-cue results (Fig. 5 and Tab. 4), we ensure the same cue strength across the 1-cue and 2-cue images.

VLMs perform at chance level for single pictorial cues and 2-cue combinations. We find a performance gap (Figs. 4 and 5) between all tested VLMs and the DepthAnythingV2 baseline in terms of depth ordering accuracy.

Fig. 5 is a cue-level heatmap of mean VLM accuracy (lower triangle) and DepthAnythingV2 baseline accuracy (upper triangle). Red-colored cells indicate better performance. The main diagonal cells (in the dotted rectangle) of the two sub-heatmaps report 1-cue accuracies, and the other cells are for 2-cue. As shown by the lower main diagonal, the VLMs perform the best on Height-in-Plane (0.54), Relative Size (0.54) and Familiar Size (0.51), although only marginally above the chance level (0.5). The largest performance gap between the VLMs and the baseline are also from the best cues of HP ($\delta = 0.45$) and RS ($\delta = 0.37$).

Table 4: Depth ordering accuracy comparing with model & image *baselines*. We run DepthAnythingV2 on all images as a reference. DepthAnythingV2 has improved performance for real images and when more cues present, whereas VLMs have a similar pattern but much weaker.

Image set	VLM mean	*DepthAnyV2*
real (mixed cues)	0.6051	0.9375
real (1-cue & 2-cue)	0.5392	0.9660
sim (1-cue & 2-cue)	0.5221	0.7090
0-cue	0.4865	0.5177
1-cue	0.5165	0.6535
2-cue	0.5282	0.7446

As one would expect, a combination of depth cues improves human depth perception [53]. For VLMs, we observe such trend but only to a limited extent, shown by the 2-cue numbers (under the main diagonal in Fig. 5) being slightly greater than the 1-cue accuracies. The depth ordering are more accurate whenever HP or RS cue is present. Most other 2-cue combinations have insignificant influence on VLMs. As summarized in Tab. 4, both VLMs and DepthAnythingV2 perform better when more cues are present.

The model-level accuracy gaps are similar to the aggregated means in Fig. 5 (see Supplementary Material). None of the evaluated VLMs perform significantly better than chance; the commercial (GPT4.1-mini & Gemini2.5) and vision-centric (Cambrian) models are no exceptions.

Linear Perspective (LP) cue interacts positively with other cues. Based on results in DepthCues [15], one of the best cues for depth perception was LP. We test it in a negative way, where the LP cue is eliminated from a view by removing ground texture [43]. In addition to confirming the claims in DepthCues [15], our results show LP has mostly positive interactions with other cues. The average accuracy gain due to LP is about 0.03, matching the best cues Height-in-Plane (HP) & Relative Size (RS).

Following [15], we apply Spearman correlation to the 8 single-cue depth ordering accuracies of the evaluated VLMs, but obtain inconsistent results (see Supplementary Material). While RS/HP had high correlation among cue tasks in both [15] (0.82) and our analysis (0.60), another highly correlated pair of RS/OC (0.75) in [15] is nearly uncorrelated (0.07) in our results. Our overall correlations among different cues for depth ordering are lower, a sign of more controlled pictorial cue analysis.

In-context learning and chain-of-thought prompting helps commercial VLMs only, with limited improvements. Among the 5 VLMs tested with ICL and CoT prompting, only the commercial GPT4.1-mini ($\delta = 0.068$) and Gemini2.5 ($\delta = 0.042$) benefit from it. The additional few-shot demonstrations hinder the performance of DeepSeek ($\delta = -0.022$), Qwen2 ($\delta = -0.012$), and Phi3.5 ($\delta = -0.006$). For all 5 VLMs, however, the performance differences are

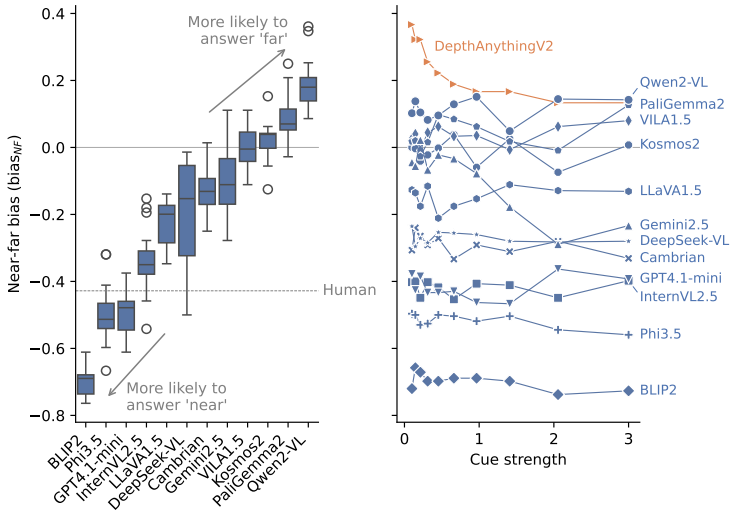


Fig. 6: Near-far bias. Both plots share the y-axis. Left: Boxes show variations across cues. VLMs exhibit different extents of biases toward answering near *vs*. far, compared with human reference [10]. Right: None of the VLMs has consistent bias reduction as cue strength increases, compared with DepthAnythingV2.

not significant across cues. The ICL and CoT prompting is most effective for the RS cue ($\delta = 0.069$) and least for FS ($\delta = -0.036$).

VLMs have a wide range of biases towards answering near *vs*. far. Most image sets in O3-D (Tab. 2) are class-balanced, meaning that the number of images with the target located near is equal to that of far. This near-far balanced property allows us to explore VLMs' biases towards near *vs*. far when answering depth ordering questions (Fig. 6).

In this paper, the near-far bias of a VLM is defined as

$$\text{bias}_{NF}(M) = \text{Acc}_F(M) - \text{Acc}_N(M), \tag{1}$$

where Acc_N and Acc_F denotes the accuracy of a model M on the target-near and target-far image subsets, respectively. We argue without proof that, given near-far balanced dataset and randomized questions (see Sec. 5.2), bias_{NF} should reflect VLMs' preferences on answering near *vs*. far. Namely, bias_{NF} should be zero when a VLM is unbiased. If a VLM always answers "far" or "near" ignoring vision inputs, bias_{NF} will be 1.0 or -1.0, respectively.

Fig. 6 reports bias_{NF} (left panel) and its change (right panel) with increased cue strengths for each benchmarked VLM. The VLMs have a wide spread of bias_{NF} ranging from -0.7 to 0.2. Majority of the VLMs prefer answering near (below the 0.0 line). BLIP2, Phi3, and GPT4.1-mini are the three most near-biased VLMs, whereas Qwen2-VL and PaliGemma2 are far-biased. The human baseline of -0.428 is computed based on [10], whose authors found a positive

correlation between center pixels and nearness for human judgments on web collected images. The baseline is comparable because we always have the target close to the image center and the distractors more eccentric.

We also analyze whether increasing cue strengths will reduce the near-far bias, and the answer is negative, as shown in the right panel of Fig. 6. Once again, the difference between VLMs and the baseline is apparent. VLM biases (shown as blue lines) do not converge towards 0 as cues become stronger. DepthAnythingV2 (orange line), in contrast, has a clear trend of bias reduction.

5.2 Language Dimension

Intuitively, if a VLM understands depth ordering in an image, the responses should remain consistent for any equivalent questions. In other words, better models should have higher language consistency and less variation. To measure this, we introduce a variation-based metric. The standard deviation of within-group means (SDGM) is given by

$$\sigma_{\Omega}(\mu) = \sqrt{\frac{1}{||\Omega||}\sum_{g\in\Omega}(\mu_g - \bar{\mu})^2}, \tag{2}$$

where Ω defines a set of groups, and μ_g denotes a mean performance metric within each group g. If, for example, Ω is the target referring clarity in Tab. 3, there will be 4 groups, and 4 within-group means $\{\mu_{g_i}\}_{i=1}^4$. Then we can obtain SDGM by computing the standard deviation of the means. For simpler analysis, we use the target-near accuracy in Eq. (1) as the metric, *i.e.* $\mu := \overline{Acc_N}$.

A good model should yield a low SDGM, the variation metric, *only when the groups in Ω are equivalent* in terms of model performance. Our question variations do not alter the essence of the depth ordering questions, so they satisfy this condition. Therefore, out of the 1,026 unique prompts, we define question groups Ω_L by tagging the prompts from 5 perspectives, ending up with 42 groups. For example, one group may have the tags: `high` clarity, `close-far` vocabulary, `formal`, `yes-no` query, and `normal` order. With a minor modification, we can use SDGM to analyze language consistency for a specific dimension of question variation, *e.g.* `normal` *vs.* `reversed` order (see Supplementary Material). Finally, the language consistency metric in Fig. 4 is given by $C = 1 - \sigma_{\Omega_L}(\overline{Acc_N})$.

We also exploit SDGM more generally, using it to represent the influence of the grouping criteria on the model performance. In other words, the SDGM of a model will be high if the model is (over-)sensitive to the differences among groups. To compare language influence with vision, we define cue groups Ω_V consisting of the 8 single-cue and 28 two-cue cases. The total 36 cue groups is comparable to the 42 question groups when computing SDGM.

Language has a much larger influence on VQA responses than vision. Fig. 7 reports the quantified vision (σ_{Ω_V}) and language (σ_{Ω_L}) influences for each VLM. The language influence is uniformly larger than that of vision. InternVL2.5 has the least blind faith in text [16] and the smallest difference between vision

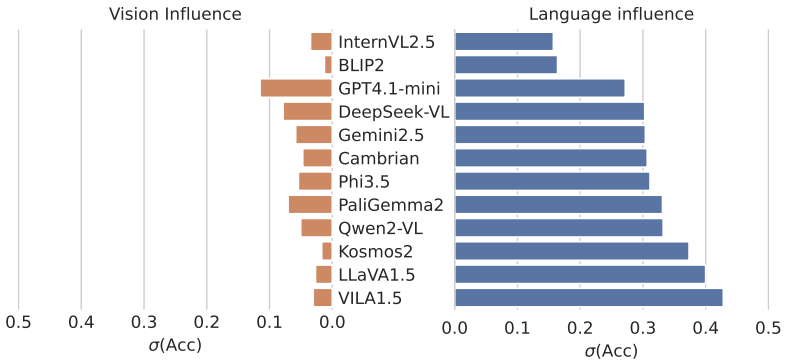


Fig. 7: Vision *vs*. language influence on depth ordering VQA. Bars show the standard deviations of mean accuracy (SDGM, see text), as a measure of influence. Language influence is uniformly larger than vision.

and language. The high language sensitivity of the VLMs is consistent with results in [9].

Inconsistencies are mainly caused by varying depth order vocabulary and `yes-no`/`multi-choice` query. Depth order vocabulary has three tags: `close-far`, `front-rear`, and `before-after`, which describes the depth relation between the target and distractors using different pairs of words. This source of variations causes the largest inconsistency ($\sigma = 0.2166$).

Whether a question is yes-no (*Is it behind?*) or multi-choice (*Is it closer or farther?*) also gives rise to a large undesired performance deviation ($\sigma = 0.1931$). InternVL2.5, as discussed above, is significantly more consistent against the yes-no query variation. Most VLMs are biased towards answering positive to the yes-no queries compared with the more neutral multi-choice queries. DeepSeek-VL and PaliGemma2 are two exceptions in that they answer "no" more often, which could be a result of over-correction (see Supplementary Material).

The `normal`/`reversed` order variation affects just a few VLMs. We tag a prompt as `reversed` order when the response options are reversed from the expression in the question (*e.g. Is it closer or farther? A. farther. B. closer.*). The most negatively impacted VLMs are VILA1.5 ($\sigma = 0.5608$) and Qwen2-VL ($\sigma = 0.3719$), followed by LLaVA1.5. This is in line with similar findings for large language models [40]. The remaining VLMs have a similar level of robustness against different option orders (see Supplementary Material).

Referring clarity slightly helps depth ordering. As referring clarity (Tab. 3) increases, the overall depth ordering accuracy slightly improves for both synthetic and real-world images. Adding markers results in the largest but still marginal improvement. Similarly, referring with bounding boxes by Kosmos2 brings a negligible improvement (see Supplementary Material). This suggests that referring expression comprehension is not a major hurdle, rather depth is.

6 Conclusion

In this work, we introduced O3-D, a dataset for testing VLMs and vision models on depth ordering. To the best of our knowledge, our work was the first systematic investigation of VLM performance across a set of pictorial cues, thanks to the novel scene construction and cue control method. Language complexity was also addressed, both independently and in combination with vision. Our experiments showed that all evaluated open-source and commercial VLMs performed around the chance level, regardless of utilized depth cues, their cue strengths, level of visual realism, and referring clarity. We quantified vision & language influences on the VQA responses, and found that the VLMs were significantly more sensitive to language input than vision. These results have implications for applications involving VLMs, such as robotic manipulation and human-robot interaction. Major future work includes extending current focus on pictorial cues to motion-based monocular cues, and further to binocular ones. We hope that the data and evaluation approach presented in this paper will help bridge the gap between language and vision components of VLMs, as well as bring machine vision closer to its biological counterpart.

Acknowledgements

This work was supported by grants to JKT from the Natural Sciences and Engineering Research Council of Canada (NSERC) under award number RGPIN-2022-04606 and the Air Force Office for Scientific Research (USA) under award number FA9550-22-1-0538.

References

1. Abdin, M., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv:2404.14219 (2024)
2. Arampatzakis, V., Pavlidis, G., Mitianoudis, N., Papamarkos, N.: Monocular depth estimation: A thorough review. IEEE TPAMI **46**(4), 2396–2414 (2023)
3. Azad, S., Jain, Y., Garg, R., Rawat, Y., Vineet, V.: Understanding depth and height perception in large visual-language models. In: CVPRW (2025)
4. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: ScanQA: 3D Question Answering for Spatial Scene Understanding. In: CVPR (2022)
5. Baldassini, F.B., Shukor, M., Cord, M., Soulier, L., Piwowarski, B.: What makes multimodal in-context learning work? In: CVPRW. pp. 1539–1550 (2024)
6. Brenner, E., Smeets, J.B.J.: Depth perception. In: Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, pp. 1–30. John Wiley & Sons, Ltd (2018)
7. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities. In: CVPR (2024)
8. Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language. In: ECCV (2020)

9. Chen, S., Gu, J., Han, Z., Ma, Y., Torr, P., Tresp, V.: Benchmarking Robustness of Adaptation Methods on Pre-trained Vision-Language Models. In: NeurIPS (2023)
10. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NeurIPS (2016)
11. Chen, Z., et al.: Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv:2412.05271 (2025)
12. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models. In: NeurIPS (2024)
13. Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V., Wang, Y.: Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In: ICLR (2025)
14. Dahou, Y., Huynh, N.D., Le-Khac, P.H., Para, W.R., Singh, A., Narayan, S.: SalBench: Vision-language models can't see the obvious. In: ICCV (2025)
15. Danier, D., Aygün, M., Li, C., Bilen, H., Mac Aodha, O.: DepthCues: Evaluating Monocular Depth Perception in Large Vision Models. In: CVPR (2025)
16. Deng, A., Cao, T., Chen, Z., Hooi, B.: Words or vision: Do vision-language models have blind faith in text? In: CVPR (2025)
17. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3d awareness of visual foundation models. In: CVPR (2024)
18. FitzGerald, N., Artzi, Y., Zettlemoyer, L.: Learning Distributions over Logical Forms for Referring Expression Generation. In: Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., Bethard, S. (eds.) Proceedings of the Conference on Empirical Methods in Natural Language Processing (2013)
19. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: BLINK: Multimodal Large Language Models Can See but Not Perceive. In: ECCV (2025)
20. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In: CVPR (2017)
21. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., Kipf, T., Kundu, A., Lagun, D., Laradji, I., Liu, H.T.D., Meyer, H., Miao, Y., Nowrouzezahrai, D., Oztireli, C., Pot, E., Radwan, N., Rebain, D., Sabour, S., Sajjadi, M.S.M., Sela, M., Sitzmann, V., Stone, A., Sun, D., Vora, S., Wang, Z., Wu, T., Yi, K.M., Zhong, F., Tagliasacchi, A.: Kubric: A Scalable Dataset Generator. In: CVPR (2022)
22. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. In: CVPR (2009)
23. Kavšek, M., Yonas, A., Granrud, C.E.: Infants' sensitivity to pictorial depth cues: A review and meta-analysis of looking studies. Infant Behavior and Development **35**(1), 109–128 (2012)
24. Kim, B.S., Kim, J., Lee, D., Jang, B.: Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges. ACM Computing Surveys **57**(10), 1–35 (2025)
25. Kotseruba, I., Wloka, C., Rasouli, A., Tsotsos, J.K.: Do Saliency Models Detect Odd-One-Out Targets? New Datasets and Evaluations. In: BMVC (2019)
26. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: SEED-Bench: Benchmarking Multimodal Large Language Models. In: CVPR (2024)
27. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: ICML (2023)

28. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: VILA: On Pretraining for Visual Language Models. In: CVPR (2024)
29. Linsley, D., Zhou, P., Ashok, A.K., Nagaraj, A., Gaonkar, G., Lewis, F.E., Pizlo, Z., Serre, T.: The 3D-PC: A benchmark for visual perspective taking in humans and machines. In: ICLR (2025)
30. Liu, C., Ding, H., Jiang, X.: GRES: Generalized Referring Expression Segmentation. In: CVPR (2023)
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: CVPR (2024)
32. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: MMBench: Is Your Multi-modal Model an All-Around Player? In: ECCV (2025)
33. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., Sun, Y., Deng, C., Xu, H., Xie, Z., Ruan, C.: DeepSeek-VL: Towards real-world vision-language understanding. arXiv:2403.05525 (2024)
34. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: SQA3D: Situated Question Answering in 3D Scenes. In: ICLR (2023)
35. Man, Y., Zheng, S., Bao, Z., Hebert, M., Gui, L.Y., Wang, Y.X.: Lexicon3D: Probing Visual Foundation Models for Complex 3D Scene Understanding. In: ICCV (2024)
36. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and Comprehension of Unambiguous Object Descriptions. In: CVPR (2016)
37. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE TPAMI **44**(7), 3523–3542 (2021)
38. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual VQA: A Cause-Effect Look at Language Bias. In: CVPR (2021)
39. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding Multimodal Large Language Models to the World. In: ICLR (2024)
40. Pezeshkpour, P., Hruschka, E.: Large language models sensitivity to the order of options in multiple-choice questions. In: Findings of the Association for Computational Linguistics: NAACL. pp. 2006–2017 (2024)
41. Qiao, Y., Deng, C., Wu, Q.: Referring expression comprehension: A survey of methods and datasets. IEEE TMM **23**, 4426–4440 (2021)
42. Qin, L., Chen, Q., Fei, H., Chen, Z., Li, M., Che, W.: What Factors Affect Multi-Modal In-Context Learning? An In-Depth Exploration. In: NeurIPS (2024)
43. Reichelt, S., Häussler, R., Fütterer, G., Leister, N.: Depth cues in human visual perception and their realization in 3D displays. In: Three-Dimensional Imaging, Visualization, and Display. vol. 7690, pp. 92–103. SPIE (2010)
44. Samavati, T., Soryani, M.: Deep learning-based 3d reconstruction: A survey. Artificial Intelligence Review **56**(9), 9175–9219 (2023)
45. Shah, M., Chen, X., Rohrbach, M., Parikh, D.: Cycle-consistency for robust visual question answering. In: CVPR (June 2019)
46. Sinapov, J., Stoytchev, A.: The odd one out task: Toward an intelligence test for robots. In: International Conference on Development and Learning. pp. 126–131 (2010)
47. Steiner, A., Pinto, A.S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., Qin, S., Ingle, R., Bugliarello, E., Kazemzadeh, S., Mesnard, T., Alabdulmohsin, I., Beyer, L., Zhai, X.: PaliGemma 2: A Family of Versatile VLMs for Transfer. arXiv:2412.03555 (2024)

48. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer Nature (2022)
49. Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S.C., Yang, J., Yang, S., Iyer, A., Pan, X., Wang, A., Fergus, R., LeCun, Y., Xie, S.: Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In: NeurIPS (2024)
50. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In: CVPR (2024)
51. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191 (2024)
52. Watt, S.J., Akeley, K., Ernst, M.O., Banks, M.S.: Focus cues affect perceived depth. Journal of Vision **5**(10), 7 (2005)
53. Westerman, S.J., Cribbin, T.: Individual differences in the use of depth cues: Implications for computer- and video-based tasks. Acta Psychologica **99**(3), 293–310 (1998)
54. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. In: NeurIPS (2024)
55. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and Ground Anything Anywhere at Any Granularity. In: ICLR (2024)
56. Zhan, G., Zheng, C., Xie, W., Zisserman, A.: A General Protocol to Probe Large Vision Models for 3D Physical Understanding. In: NeurIPS (2024)
57. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE **111**(3), 257–276 (2023)