BLINK: Multimodal Large Language Models Can See but Not Perceive

Xingyu Fu[1*], Yushi Hu[2,3*], Bangzheng Li[4], Yu Feng[1], Haoyu Wang[1], Xudong Lin[5], Dan Roth[1], Noah A. Smith[2,3], Wei-Chiu Ma[3†], and Ranjay Krishna[2,3†]

[1] University of Pennsylvania [2] University of Washington [3] Allen Institute for AI
[4] University of California, Davis [5] Columbia University
https://zeyofu.github.io/blink/

Figure 1: The BLINK Benchmark. BLINK contains 14 visual perception tasks that can be solved by humans "within a blink", but pose significant challenges for current multimodal LLMs. These tasks are inspired by classical computer vision problems and recast into multiple-choice questions for multimodal LLMs to answer. Notice that the visual prompts and questions in this figure are different from the actual ones used in the benchmark for illustrative purposes, and answers of the samples are provided in fn.6.

Abstract. We introduce BLINK, a new benchmark for multimodal language models (LLMs) that focuses on core visual perception abilities not found in other evaluations. Most of the BLINK tasks can be solved by humans "within a blink" (*e.g.*, relative depth estimation, visual correspondence, forensics detection, and multi-view reasoning). However, we find these perception-demanding tasks cast significant challenges for current multimodal LLMs because they resist mediation through natural language. BLINK reformats 14 classic computer vision tasks into 3,807 multiple-choice questions, paired with single or multiple images and visual prompting. While humans get 95.70% accuracy on average, BLINK is surprisingly challenging for existing multimodal LLMs: even the

best-performing GPT-4V and Gemini achieve accuracies of 51.26% and 45.72%, only 13.17% and 7.63% higher than random guessing, indicating that such perception abilities have not "emerged" yet in recent multimodal LLMs. Our analysis also highlights that specialist CV models could solve these problems much better, suggesting potential pathways for future improvements. We believe BLINK will stimulate the community to help multimodal LLMs catch up with human-level visual perception.

Keywords: Multi-modal Large Language Models · Vision-Language Benchmark · Visual Perception Evaluation

1 Introduction

Compared to today, computer vision was originally attempting to interpret images as projections of 3D scenes, not just processing 2D arrays of flat "patterns" [25, 59, 62]. In this pursuit, early research developed a series of intermediate tasks: they focused on understanding optical properties like reflectance [12, 79], 3D primitives through multi-view reasoning [38, 60], geometric reasoning through depth estimation [76], instance recognition through visual correspondence [56], affordance through keypoint grounding [37], and forensics through intrinsic images [9]. Yet in the modern era of large language models (LLMs), we, as a community, have focused less on such perceptual tasks, and instead have developed new tasks, mostly expressed in natural language, emphasizing the vision-language connection learned by multimodal LLMs [3, 6, 18, 20, 24, 26, 51, 52, 57, 63, 66, 73, 80]. This might be because many traditional computer vision tasks resist mediation through natural language, due to the inherent imprecision of language (*e.g.*, it is challenging to precisely pinpoint a spatial keypoint through language).

This paper aims to highlight crucial aspects of visual perception that have been overlooked when evaluating multimodal LLMs. To appropriately position our paper, let us revisit how we currently evaluate perception through using multimodal LLMs [44, 45, 48, 53, 54, 58, 90]. While many of these benchmarks have been popularized as the de facto evaluation measures for influential models like GPT-4V and Gemini-Pro, they conflate perception with language knowledge and reasoning. At the risk of singling out one benchmark, let us consider two questions highlighted in the popular MMBench [53]: "`<image 1> Why is this hummingbird called ruby-throated?`" and "`<image 1> What will happen next? A: the person is gonna laugh B: the person is gonna cry.`" For the first question, the vision subpart is to recognize the hummingbird. For the second, it only needs a coarse description of the image. Everything else is left to the language model to solve. Such a conflation has also been reported for other benchmarks by previous work [11, 39, 87]. Our experiments show that this conflation reductively evaluates perception as a dense captioning task. In other words, **by replacing the image with a task-agnostic dense caption, our experiments show that a "blind" GPT-4 performs well on these "multimodal tasks".**

In response, we propose BLINK. BLINK reimagines traditional computer vision problems through a format that allows us to evaluate multimodal LLMs.

Figure 2: Comparison between BLINK and previous benchmarks. BLINK has several novel features: (1) BLINK incorporates diverse visual prompts, like circles, boxes, and image masks, while previous benchmarks only have text questions and answers. (2) BLINK evaluates a more comprehensive range of visual perception abilities, like multi-view reasoning, depth estimation, and reflectance estimation. Prior benchmarks are generally more focused on recognition-based VQA. (3) BLINK contains "visual" commonsense problems that humans can answer within seconds, while prior benchmarks like [90] require domain knowledge. The samples of previous benchmarks are from [45, 53, 90]. Part of our samples are curated from [10, 19, 29, 32, 36, 43, 91].

As partially demonstrated in Figure 1[6], BLINK consists of 14 classic computer vision tasks, ranging from low-level pattern matching (*e.g.*, visual correspondences estimation) to mid-level reasoning (*e.g.*, relative depth estimation), and extending to high-level visual understanding (*e.g.*, visual similarity). The image tasks are meticulously selected such that they are difficult to solve by reducing the evaluation using dense captioning; instead, the models must perceive the contents of the image(s) to answer. We recast each traditional task into a modern question-answering format, where answer choices are either images or text. BLINK contains 3.8K questions across 7.3K images, where questions may contain multiple images that are curated from a wide range of datasets [8, 10, 19, 29, 32, 36, 42, 47], encompassing indoor household scenes as well as outdoor urban or natural environments. The questions and choices are either derived from the datasets, or

[6] The answers of the examples in Figure 1 are as follows. Relative depth: B; jigsaw: A; multi-view reasoning: right; visual correspondence: A; semantic correspondence: C; forensics detection: final image; IQ test: D; visual similarity: upper one; functional correspondence: A; relative reflectance: they are about the same.

manually written by humans. On average, each question can be solved by a human subject within a BLINK of an eye, except the IQ test.

We carefully evaluate 17 multimodal LLMs with various sizes (*i.e.*, 7B, 13B, 34B) on BLINK. We observe the paradox that **while these problems are easy for humans (**95.70% **average accuracy), they are extremely hard for existing machinery** – even GPT-4V model can only achieve 51.26% accuracy on average, which is 44.44% worse than humans, and 13.17% better than random guessing. We also experiment with specialist vision models and find that they perform much better than multimodal LLMs. For example, the specialist outperforms GPT-4V by 62.8% on visual correspondence estimation, 38.7% on relative depth estimation, and 34.6% on multi-view reasoning, in terms of absolute accuracy. Our findings indicate that the perceptual abilities of multimodal LLMs have been previously overestimated. Furthermore, these models may benefit from integrating insights from specialized models that excel in these areas. We believe BLINK can serve as an effective testbed for bridging the gap between traditional notions of perception and the modern generative capabilities of multimodal LLMs.

2 Related Work

Multimodal models: Inspired by the impressive success in recent large language models (LLMs) [13, 21, 63, 77, 93], a sequence of studies explore multimodal LMMs that can jointly understand vision and language information and generate textual answers through adding a modality adaption structure between a frozen visual encoder [27, 65, 71] and a frozen LLM [77, 93]. Flamingo [3] and BLIP-2 [46] are two of the earliest works to explore these transformer-based multi-modality conjunction structures. They first pre-train on image-text matching datasets [15, 42, 47, 67] and then fine-tune on task-specific datasets such as visual question answer (VQA) [4, 34]. Starting from LLaVA [49, 51], people use LLM synthesized instruction-following chat data (which are in VQA format) for instruction tuning and achieve much better results [7, 16, 24, 73]. There have been extended studies that explore further capabilities of multimodal LLMs, especially on VQA reasoning [30–33, 39, 40, 68, 78, 84, 92]. However, they mainly focus on the textual reasoning abilities [83] within the multimodal LLMs and do not emphasize visual perceptions.

Multimodal benchmarks: Traditional vision-language datasets are designed to assess single-task capabilities, such as optical character recognition (OCR) [55], image captioning [47], and visual question answering [4, 34]. However, these datasets are often not comprehensive enough to holistically assess multimodal LMMs on general perception and reasoning abilities. Many recent papers have built more comprehensive benchmarks. MME [28] is one of the earliest holistic benchmarks containing multi-modal Yes/No questions on the defined visual perception and language reasoning tasks. MM-Vet [89] includes six sub-features from the previous datasets including recognition-focused questions, OCR, and math, providing a diverse while discrete evaluation set. MMBench [53] covers more subjects and provides a more robust circular evaluation setting. Seed-

Bench [44,45] benchmark has a more diverse source of inputs, including multiple-image inputs and video, and includes more tasks. However, the visual perception questions in MME, MMBench, MM-Vet, and Seed-Bench are mainly extracted from existing VQA datasets or generated by GPT [63] from image descriptions such as COCO-Caption [47], and are recognition focused, covering topics such as object (attribute)recognition, and OCR. In contrast, we focus on multiple distinct nuanced perception abilities and recognition-level perception is only one of our focus. Some other multimodal benchmarks have distinct focuses. MMMU [90] aims at achieving expert-level artificial general intelligence by collecting domain-knowledge-required questions. HallusionBench [35] mainly tests the language hallucination and visual illusion phenomena. MathVista [58] presents exclusively mathematical domain visual questions based on images such as charts, tables, and diagrams. These benchmarks do not require human-level perception abilities as in BLINK and therefore cannot measure model visual perceptions holistically.

3 The BLINK Benchmark

Our goal is to faithfully evaluate the visual perception capabilities of existing Multimodal LLMs. We seek to study the visual perception gap between humans and machineries, and offer deeper insights into potential pathways towards achieving more generalized machine perception. Based on the observation that existing benchmarks predominantly focus on evaluating visual recognition abilities, we introduce a novel benchmark, BLINK, designed to enable both quantitative and qualitative evaluation of the nuanced perception capabilities of multimodal LLM across various dimensions. We unfold this section by illustrating the overall design of BLINK (§3.1) and discussing its unique features comparing with previous benchmarks. Then we describe each task in detail, providing an in-depth explanation of the data curation process (§3.2).

3.1 Overview of BLINK

To ensure that one can effectively measure what Multimodal LLMs can or cannot perceive, we carefully select 14 tasks (see §3.2 for the full list) that are difficult to solve by reducing the evaluation into text-only questions using dense captioning. The tasks are drawn from either classic computer vision problems or recent applications of Multimodal LLMs, each of which requires a nuanced understanding of the visual data. They range from low-level pattern matching (*e.g.*, visual correspondence) to mid-level spatial reasoning (*e.g.*, relative depth), and up to high-level visual understanding (*e.g.*, visual similarity). This variety allows for a systematic exploration of Multimodal LLMs' capabilities across different perceptual complexity layers. Furthermore, these visual tasks vary in granularity, ranging from pixels (*e.g.*, relative reflectance) to patches (*e.g.*, jigsaw) and extending to the full image (*e.g.*, forensic detection), enabling us to evaluate models' proficiency in observing at various scales.

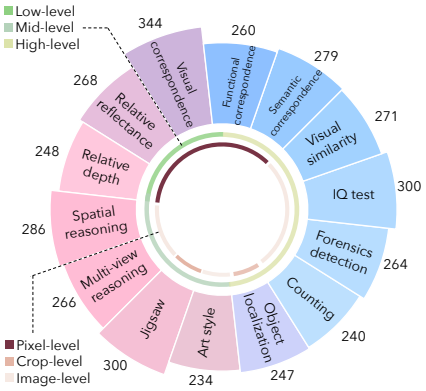

Figure 3: Statistics of BLINK. 14 tasks range from pixel-level to image-level; from low-level pattern matching (*e.g.*, visual correspondences) to mid-level reasoning (*e.g.*, relative depth), further to high-level visual understanding (*e.g.*, visual similarity).

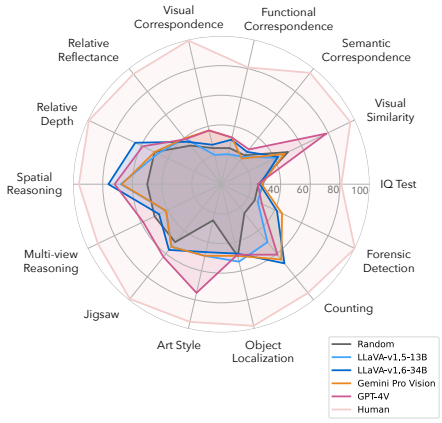

Figure 4: Accuracies of multimodal LLMs on BLINK test set. Please refer to Table 1 and §4.2 for more results and discussions.

To facilitate the evaluation of multimodal LLMs, we recast all tasks as multiple-choice question-answering problems. The options for answers may include images or texts, while the questions themselves can feature either single or multiple images. Prompts are designed to be both textual and visual in nature. We re-purposed several existing vision datasets as well as collected new data. In total, we contribute 3.9K multiple-choice questions and 7.3K images, with an even distribution between the validation and test sets. Numbers of each task are reported in Figure 3, and more detailed statistics can be found in Appendix A.5.

Key features of BLINK: Comparing with previous benchmarks, BLINK has the following novel features:

- **Visual prompting**: Unlike existing benchmarks that support only text prompting, BLINK features a variety of visual prompts. This enables one to highlight specific areas within images, facilitating the evaluation of Multimodal LLMs' detailed understanding of these regions. It also offers an interface for researchers to investigate the impact of visual prompting techniques.
- **Perception beyond recognition**: Besides visual recognition, BLINK considers a diverse set of visual perception abilities, such as 3D reasoning, geometric understanding, affordance reasoning, etc. The breadth allows one to evaluate Multimodal LLMs from an unique array of perspectives.
- **"Visual commonsense" that does not require domain knowledge**: The questions in BLINK are intentionally designed to be straightforward, requiring neither domain-specific knowledge nor expertise to answer. They are crafted in such a way that humans can solve them almost instantaneously,

typically within a few seconds. This allows us to explore the fundamental gap in visual perception gap between humans and Multimodal LLMs, highlighting the paradox that problems easily solved by humans often pose significant challenges for machines.
- **Interleaved image-text formats**: BLINK features a heterogeneous question-answering format, wherein both questions and choices can be presented as text or images. This diversity compels Multimodal LLMs to genuinely understand the questions, pushing the boundaries of their interpretative capabilities.
- **Diverse image sources**: BLINK comprises a wide range of in-the-wild images sourced from various origins, covering everything from indoor and outdoor scenes to object-centric views and landscapes. This collection spans abstract diagrams, synthesized images, and authentic photographs, ensuring a comprehensive examination of visual perception

The design principles of BLINK are also illustrated in Figure 2. We will now describe each task in detail.

3.2 Dataset Collection Process

BLINK comprises 14 tasks, all of which have been repurposed into a multiple-choice question-answering format. These tasks utilize a diverse collection of images from various sources and we ensure that each image is unique.

Visual correspondence: This task aims to evaluate the ability of Multimodal LLMs to understand and identify the same scene point across various viewpoints, lighting conditions, or time. We exploit HPatches [8] for this task. HPatches contains a number of image sequences, each of which are composed of images taken under different illuminations and/or viewpoints of a scene. For each question, we randomly sample two images and an interest point within them. Then we exploit the ground-truth homography to compute its correspondence. Finally, we randomly select three more interest points to serves as other choices.

Relative reflectance: This task aims to compare the reflectance (albedo) of two pixels. It allows us to evaluate Multimodal LLMs' understanding of material properties and their interaction with light, which is crucial for applications requiring high-fidelity visual interpretations. We curate our samples using human annotations from the Intrinsic Images in the Wild (IIW) dataset [10]. Each question is based on an image and two specified points. The goal is to identify which point is darker, or whether the two points have similar reflectance.

Relative depth: Humans are good at judging relative depth [19]. This task can thus serve as a proxy to validate if the geometric understanding capabilities of existing multimodal LLMs are close to human. We curate our samples using human annotations from the Depth in the Wild [19] dataset. Given an image and two specified points, the task is to determine which point is closer.

Spatial relation: Understanding spatial relationships between objects in a scene is essential for interpreting complex visual environments. However, modern Multimodal LLMs often struggles with spatial concepts such as "left" and "right" [88]. This task help us evaluate whether the models finally possess this

Figure 5: Qualitative results on BLINK. For each task, we show the choice of LLaVA-v1.6-34B [51], Qwen-VL-Max [7], Gemini Pro [73], GPT-4V [63], and humans. Red choice indicates the ground truth. Notice that the markers are intentionally enlarged for visualization purposes, and we make some images inset images to save space. For IQ test, the third image is constructed by overlaying the first and second images.

vital skill. We curate our samples from the Visual Spatial Reasoning [48] dataset. Each sample contains an image and a claim. The task is to determine if the claim is true or false. We reformat the claims into binary questions via GPT-3.5 [13].

Multi-view reasoning: This task is centered on evaluating the multi-view reasoning capabilities of Multimodal LLMs. The objective is to deduce the

relative camera motion based on two images of an object captured from different viewpoints. Our data is sourced from the Wild6D dataset [91], which features videos of various objects recorded in diverse settings. We select two random frames from each video to calculate the relative camera motion. Recognizing that even humans might struggle to precisely articulate 3D motion details, we simplify the task by classifying motions into two broad categories: moving towards the left or moving towards the right. Despite the simplicity of these questions, as we will later demonstrate, they pose significant challenges for current models.

Jigsaw: This task assesses the ability of Multimodal LLMs to recognize and group patterns, aligning patches based on continuity in shape, color, and texture. We utilize images from the TARA dataset [32] and segment each of them into a 3x3 grid. We retain the three segments from the upper left corner as the reference image, and treat the central segment along with a randomly chosen segment as options. The objective is to identify the correct patch (*i.e.*, the central patch).

Art style: This task evaluates Multimodal LLMs capability to analyze and discern both local and global similarities in art styles among multiple images. Although there have been prior efforts to incorporate art-related questions into evaluation [90], such attempts primarily focused on questions requiring expert-level knowledge, including deducing an artist's name and understanding historical contexts, rather than on direct image comparison. For this task, we collect paintings and their stylistic information from WikiArt. Given one reference painting image and two other paintings as options, the model is tasked with identifying the one that most closely shares the art style of the reference painting.

Object localization: The ability to accurately detect and localize objects is critical for scene understanding. While previous benchmarks [53] have explored this task, their focus was primarily on coarse localization. For instance, they might only ask the model if an object is located at the "top" or "right" side of an image. BLINK, in contrast, aims for a more fine-grained evaluation. We exploit images from LVIS [36], randomly sampling one object per image along with its ground-truth bounding box. Then we add Gaussian noise to the ground-truth box to create a confounding box. The goal is to select the correct one.

Counting: This task evaluate Multimodal LLMs' abilities in detection, recognition, and compositional reasoning, particularly in complex scenes where objects may overlap, be occluded, or vary in size and appearance. We select our questions from the TallyQA dataset [2], known for its challenging human-written counting questions. Each sample comprises an image, a question, and a numerical answer. In addition, we randomly select three numbers to serve as confounding options.

Forensic detection: Recent advances in generative AI have raised concerns about malicious uses and have prompted calls for the automatic detection of fake content. To evaluate whether Multimodal LLMs can fulfill such a role, we construct sets of real and synthesized images that describe similar scenes and ask the models to identify the real ones. Specifically, we first generate synthetic images using Stable Diffusion XL [64], employing COCO captions [47] as prompts. Then, we manually search online using these captions as descriptions and select high-quality photographs as the real images.

IQ test: This task evaluates the ability of Multimodal LLMs to engage in graphical reasoning, without requiring any domain-specific knowledge. We manually collect test samples, along with human explanations, from various public, license-friendly online sources. Given visual examples and a selection of images, the objective is to identify the image that either continues the pattern established by the examples or is spatially consistent with them.
Visual similarity: This task aims to verify whether Multimodal LLMs possess a nuanced understanding of visual features, patterns, and aesthetics at a level comparable to humans. We select our samples from the DreamSim dataset [29]. Given a reference image alongside two alternative images, the objective is to identify the image that more closely resembles the reference image perceptually.
Semantic correspondence: This task focuses on identifying and matching semantically similar yet visually distinct elements across images, thereby evaluating the ability of Multimodal LLMs to understand the underlying semantics of object parts. Our samples are sourced from the SPair-71k dataset [61], which features pairs of images with multiple corresponding semantic points. For each task, we randomly select one semantic point in an image as a reference, and provide the matching point alongside three random semantic points in the paired image as options. The objective is to accurately identify the correct matches.
Functional correspondence: The task aims to identify points that are functionally similar across objects. It challenges Multimodal LLMs to extend their understanding beyond mere semantics, enabling them to infer the diverse functions an object can perform in various contexts. Such capability is crucial for applications in robotics. We derive our samples from the FunKPoint dataset [43], which features paired images annotated for functional correspondences. Following a method analogous to semantic correspondence, we present an action alongside two object images. One image includes a reference point, while the other offers four potential points. The objective is to select the point that best matches the reference in terms of functional affordances.
Data quality control: To guarantee the quality of BLINK, we manually go through all collected data and filter out data that are ambiguous.

4 Experiments

In this section, we first describe the experimental setup and the baselines (§4.1). Then we present a comprehensive evaluation of 16 recent multimodal LLMs (§4.2). We demonstrate that while humans can answer the questions with high accuracy, BLINK is challenging for existing models. Finally, we provide detailed analyses on multiple experimental settings, including the effect of reducing images to captions, sensitivity to different visual prompts, and error analysis (§4.3).

4.1 Experimental Setup

Multimodal LLMs: We evaluate BLINK on 16 recent models: MiniGPT-4-v2 [16], OpenFlamingo-v2 [5], InstructBLIP (7B and 13B) [24], CogVLM [80],

LLaVA(v1, v1.5, v1.6, internLM, and xtuner versions, model size 7B, 13B, and 34B) [23,26,50–52], Yi-VL (6B and 34B)[7], Qwen-VL-MAX [7], Gemini Pro [73], Claude 3 Opus [1] and GPT-4V(vision) [63]. See Appendix B for more details.
Evaluation setup: We follow standard setups as in the VLMEvalKit [22], where the temperature is set to 0 and retry is set to 10. However, we do not resize the images during any experiment. For the models that do not support multiple images as input, we concatenate the images as input. We extract the choice from the models' output with a set of pre-defined rules and GPT-3.5-turbo [13]. We refer the readers to Appendix A for more details on visual prompting, how we generate the answers in BLINK, and the human evaluation protocol.

4.2 Main Results

Overall performance: As shown in Table 1, the mean accuracy of 7B and 13B open-source Multimodal LLMs hover around 35–42%, which is similar to random guess (38.09%). The most proficient open-source model, LLaVA-v1.6-34B, achieves an accuracy of 45.05%. Even the most advanced models, GPT-4V and Gemini Pro and Claude 3 OPUS, achieve accuracies of only 51.26%, 45.72%, and 44.11% respectively. Their performance are merely 13.17%, 7.63% and 6.02% better than random guessing and lag behind human performance by 44.44%, 49.98% and 51.59%. Notably, for certain tasks such as jigsaw, semantic correspondence, multi-view reasoning, object localization, and relative reflectance, some multimodal LLMs even underperform compared to random guessing. Some qualitative results are shown in Figure 5. Detailed scores on the validation set are in Appendix C.
In which tasks do multimodal LLMs show relative strengths and weaknesses? Figure 4 shows the accuracies of the best-performing models on BLINK: LLaVA-v1.6-34B [51], Gemini Pro [73], and GPT-4V [63]. We observe that multimodal LLMs perform relatively better on spatial reasoning, art style, and counting tasks, in which they are much better than random guessing. The models also demonstrate some capability in relative depth and forensics detection. Overall, they are doing relatively well on mid-level perception tasks. In terms of granularity, the models in general perform better on image-level tasks and struggle on pixel-level and crop-level tasks.
GPT-4V behaves differently: Figure 4 and Table 1 show an interesting phenomenon: GPT-4V's performance pattern is different from other models. Compared with its counterparts, GPT-4V is much better in visual similarity, art style, jigsaw, and multi-view reasoning. Specifically, its performance on visual similarity is 29% better than Gemini Pro, demonstrating that GPT-4V possesses a nuanced understanding of visual patterns and aesthetics that is similar to humans. In contrast, Gemini Pro and LLaVA have similar performance patterns.
Human performance: Human evaluators achieve over 95% accuracy across most tasks, with an average accuracy of 95.70% .[8] This performance disparity between

[7] More details are at the official website at `https://www.01.ai/`
[8] Note that the human score for IQ test is annotated by authors. It may not reflect typical human performance, which is also expected to vary.

	Validation (1,901)	Test (1,906)	Similarity (136)	Counting (120)	Depth (124)	Jigsaw (150)	Art (117)	Fun.Corr. (130)
Random Choice	38.09	38.09	50	25	50	50	50	25
Human	95.67	95.70	96.70	93.75	99.19	99.00	95.30	80.77
Open-source multimodal LLMs								
MiniGPT-4-v2 [16]	34.23	34.57	52.94	10.83	49.19	26.00	47.86	18.46
OpenFlamingo-v2 [5]	39.18	38.32	55.15	21.67	54.03	46.00	52.14	36.15
InstructBLIP-7B [24]	39.72	38.65	46.32	29.17	50.81	54.00	47.86	23.85
InstructBLIP-13B [24]	42.24	39.58	46.32	30.83	50.00	54.00	50.43	22.31
LLaVA-internLM2-7B [74]	37.71	36.06	52.94	52.50	52.42	34.67	30.77	23.08
Yi-VL-6B [7]	38.72	41.24	46.32	46.67	56.45	50.00	53.85	23.85
Yi-VL-34B [7]	41.68	42.78	50.00	58.33	53.23	54.00	46.15	**39.23**
LLaVA-v1.5-7B-xtuner [23]	39.36	40.81	46.32	53.33	50.81	54.00	47.86	23.85
LLaVA-v1.5-13B-xtuner [23]	42.00	41.31	46.32	45.00	54.03	53.33	47.86	26.15
CogVLM [80]	41.54	39.38	46.32	38.33	50.81	52.67	49.57	23.85
LLaVA-v1.5-7B [49]	37.13	38.01	46.32	43.33	51.61	11.33	47.86	21.54
LLaVA-v1.5-13B [49]	42.66	40.55	46.32	50.00	47.58	54.00	47.86	20.77
LLaVA-v1.6-34B [51]	46.80	45.05	46.32	**68.33**	64.52	56.67	47.01	30.77
API-based models								
Qwen-VL-Max [7]	40.28	41.94	51.47	55.83	58.87	3.33	37.61	28.46
Gemini Pro [73]	45.16	45.72	55.88	65.00	50.00	54.00	49.57	32.31
Claude 3 OPUS [1]	44.05	44.11	70.59	49.17	57.26	32.67	60.68	22.31
GPT-4V(ision) [63]	51.14	51.26	**83.09**	60.83	58.87	62.67	78.63	31.54
GPT-4 Turbo [63]	54.61	53.89	**83.09**	60.83	**66.94**	**66.00**	81.20	31.54
GPT-4o [63]	**60.04**	**59.03**	65.44	51.67	64.52	58.00	**82.91**	**39.23**

	Sem.Corr. (140)	Spatial (143)	Local. (125)	Vis.Corr. (172)	Multi-view (133)	Reflect. (134)	Forensic (132)	IQ (150)
Random Choice	25	50	50	25	50	33.33	25	25
Human	96.07	98.25	98.00	99.42	92.48	95.14	100.00	80.00
Open-source multimodal LLMs								
MiniGPT-4-v2 [16]	26.43	51.75	**56.00**	23.84	52.63	31.34	17.42	19.33
OpenFlamingo-v2 [5]	23.57	46.85	52.00	25.00	41.35	43.28	15.91	23.33
InstructBLIP-7B [24]	25.00	55.24	44.80	22.67	**58.65**	29.85	29.55	23.33
InstructBLIP-13B [24]	22.86	64.34	52.00	20.93	54.14	46.27	13.64	26.00
LLaVA-internLM2-7B [74]	22.14	74.13	48.00	21.51	41.35	32.84	3.79	14.67
Yi-VL-6B [7]	26.43	72.73	49.60	29.65	48.12	29.85	20.45	23.33
Yi-VL-34B [7]	21.43	70.63	54.40	23.84	41.35	46.27	17.42	22.67
LLaVA-v1.5-7B-xtuner [23]	24.29	74.83	45.60	23.84	42.11	26.87	36.36	21.33
LLaVA-v1.5-13B-xtuner [23]	22.14	**77.62**	48.00	22.09	41.35	46.27	29.55	18.67
CogVLM [80]	23.57	67.13	43.20	20.93	57.14	26.87	24.24	26.67
LLaVA-v1.5-7B [49]	32.14	70.63	48.80	20.35	49.62	36.57	28.03	24.00
LLaVA-v1.5-13B [49]	23.57	67.83	47.20	20.35	41.35	45.52	27.27	28.00
LLaVA-v1.6-34B [51]	27.86	76.22	41.60	27.33	46.62	29.85	41.67	26.00
API-based models								
Qwen-VL-Max [7]	29.29	**77.62**	49.60	22.67	53.38	**49.25**	47.73	22.00
Gemini Pro [73]	22.14	67.13	46.40	37.21	41.35	46.27	45.45	27.33
Claude 3 OPUS [1]	20.71	57.34	46.40	31.40	57.89	27.61	62.12	21.33
GPT-4V(ision) [63]	30.00	72.03	50.40	37.21	58.65	38.81	30.30	24.67
GPT-4 Turbo [63]	32.86	67.13	48.80	42.44	57.14	34.33	51.52	**30.67**
GPT-4o [63]	**45.71**	76.92	**56.00**	**71.51**	**60.15**	38.81	**85.61**	**30.00**

Table 1: Results of different models on the BLINK test set. The first row shows task names and number of test data. Average scores on validation set are also included.

humans and multimodal LLMs highlights the significant visual perception gap that exists between current machine learning models and humans in perceiving, processing, and understanding complex visual and textual context.

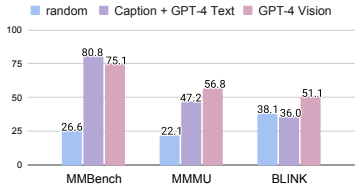

Figure 6: Performance of using image caption + text-only GPT-4 *vs.* GPT-4 Vision on MMBench [53], MMMU [90], and BLINK (§4.3).

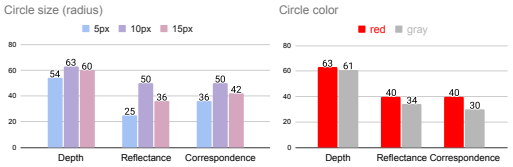

Figure 7: Accuracy of GPT-4V with different visual prompts (*e.g.*, different circle sizes, colors) on relative depth, relative reflectance, and visual correspondence tasks. More discussions in §4.3.

4.3 Analysis

Is dense captioning all you need for a multimodal LLM benchmark? To answer the question, we reduce multimodal benchmarks to a text-only problem. Specifically, we convert images into task-agnostic dense image captions with GPT-4V. The dense caption describes detailed information about the image and the visual prompts (*e.g.*, where each circle is), using language. For each multimodal question, we prompt the text-only GPT-4-0125-preview model with image captions and the textual question and evaluate if the "blind" GPT-4 can answer the question. We call this `Caption + LLM`. This experiment is predicated on the hypothesis that captioning involves predominantly recognition-centric perception. If using captions along with text-only LLMs yields performance comparable to or surpassing that achieved through the integration of images with multimodal LLMs, then the perception demands of that benchmark are primarily confined to recognition only.

We experiment with BLINK, MMBench [53] and MMMU [90], as illustrated in Figure 6. Surprisingly, we find that the `Caption + LLM` setting achieves better results on MMBench than GPT-4V (with 5.7% increase in accuracy). On MMMU, `Caption + LLM` achieves 47.2% accuracy, which is 9.6% lower than GPT-4V performance, but is still much better than random guessing. On BLINK, `Caption + LLM` fails, achieving random guessing performance. These results indicate that dense captions cover the visual information needed for MMBench. For MMMU, image captions carry a large portion of visual information needed to answer the domain-knowledge-specific questions. Meanwhile, the performance decrease observed in BLINK suggests the necessity for advanced perceptual abilities beyond what is currently attainable with general captions. This variance highlights the limitations of existing multimodal LLM benchmarks in addressing the full spectrum of visual perception.

Effect of visual prompting on BLINK: Several BLINK tasks involve visual prompting. Prior work [69] shows that factors like shape, size, and color may affect task performance, and circles give the best overall performance. Following [69], we adopt circles in BLINK and analyze the effect of circle sizes and colors on multiple tasks in Figure 7. We experiment with relative depth, relative reflectance,

Task	Vis.Corr.	Depth	Multi-view	Sem.Corr.	Forensic	Reflect.
Random	25.00	50.00	50.00	25.00	25.00	33.33
Human	99.56	99.59	92.10	94.60	100.00	99.63
Gemini Pro	42.44	40.32	44.36	26.62	50.76	45.52
GPT-4V	33.72	59.68	55.64	28.78	34.09	38.81
Specialist	DIFT [72]	DepthAnything [86]	LoFTR [70]	DIFT [72]	DIRE [82]	Ordinal Shading [14]
	96.51	97.58	90.22	71.22	68.94	77.61

Table 2: Comparison between multimodal LLMs, specialists, and human performance on the BLINK dev set. The specialists perform much better than multimodal LLMs.

and visual correspondence, with 100 validation set samples per task. The images are all reshaped to 1024px height. We experiment with circles with 5px, 10px, and 15px radius, and with red or gray color. We find that red is better than gray for all tasks. Also, the optimal circle size is task-dependent. On average 10px circles work the best, and we use it for all evaluations in this paper. The experiments suggest that visual prompting can have a big impact on multimodal LLM performance, and improving visual prompts or improving model robustness to prompt variation is a promising direction for future research [85].

Can specialist models solve BLINK tasks? Specialists can serve as a proxy upper bound of how good multimodal LLMs could be. We download the trained checkpoints for six specialist models and evaluate them on BLINK. As shown in Table 2, the specialists perform much better than GPT-4V and Gemini Pro, outperforming the best multimodal LLM by 18% to 57% on these tasks. Specifically, DepthAnything [86] and DIFT [72] achieve human-level performance on depth estimation and visual correspondence, whereas multimodal LLMs fail miserably. This sheds light on the possibility that multimodal LLMs may progress on these tasks given the correct data and training strategy. For instance, one possible way is to distill existing specialist models into multimodal LLMs [41].

Error analysis of GPT-4V: We randomly sampled 140 error instances made by GPT-4V on BLINK, 10 per task, and meticulously examined them. The most common types of errors are: **Hallucinate fine-grained patterns and attributes** (24.2%): the model hallucinates the nuanced details of objects. This error is most common for relative reflectance, forensics detection, and jigsaw tasks. **Hallucinate visual prompt locations** (20.0%): the circle location described by the model is wrong. This is common for visual correspondence and relative depth tasks. Other errors include Failures on capturing overall setting or style (8.6%), and Failures on grounding an object (5.7%). More details are in Appendix C.3.

5 Conclusion

We introduced BLINK, a new multimodal LLM benchmark that evaluates core visual perception abilities not found in existing evaluations. While these tasks seem trivial for humans to solve "within a blink", we find they pose significant challenges for current multimodal LLMs. Ultimately, Blink provides an effective testbed for multimodal LLMs to catch up with human-level visual perception.

Acknowledgements

This work was funded in part by ONR Contract N00014-23-1-2417, and supported by NSF grant IIS-2212433.

References

1. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family (March 2024)
2. Acharya, M., Kafle, K., Kanan, C.: Tallyqa: Answering complex counting questions. In: AAAI (2019)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
5. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
6. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)
7. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
8. Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K.: Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In: CVPR (2017)
9. Barrow, H., Tenenbaum, J., Hanson, A., Riseman, E.: Recovering intrinsic scene characteristics. Comput. vis. syst **2**(3-26), 2 (1978)
10. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Trans. on Graphics (SIGGRAPH) **33**(4) (2014)
11. Berrios, W., Mittal, G., Thrush, T., Kiela, D., Singh, A.: Towards language models that can see: Computer vision through the lens of natural language. arXiv preprint arXiv:2306.16410 (2023)
12. Black, M.J., Anandan, P.: A framework for the robust estimation of optical flow. In: 1993 (4th) International Conference on Computer Vision. pp. 231–236. IEEE (1993)
13. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
14. Careaga, C., Aksoy, Y.: Intrinsic image decomposition via ordinal shading. ACM Trans. Graph. (2023)
15. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR (2021)

16. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning (2023)
17. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
18. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023)
19. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. Advances in neural information processing systems **29** (2016)
20. Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C.R., Goodman, S., Wang, X., Tay, Y., Shakeri, S., Dehghani, M., Salz, D., Lucic, M., Tschannen, M., Nagrani, A., Hu, H., Joshi, M., Pang, B., Montgomery, C., Pietrzyk, P., Ritter, M., Piergiovanni, A., Minderer, M., Pavetic, F., Waters, A., Li, G., Alabdulmohsin, I., Beyer, L., Amelot, J., Lee, K., Steiner, A.P., Li, Y., Keysers, D., Arnab, A., Xu, Y., Rong, K., Kolesnikov, A., Seyedhosseini, M., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: Pali-x: On scaling up a multilingual vision and language model (2023)
21. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: Palm: Scaling language modeling with pathways (2022)
22. Contributors, O.: Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass (2023)
23. Contributors, X.: Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner (2023)
24. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
25. DO CT OR OF, P.E.: MACHINE PERCEPTION OF THREE-DIMENSIONAL, SO LIDS. Ph.D. thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY (1961)
26. Dong, X., Zhang, P., Zang, Y., Cao, Y., Wang, B., Ouyang, L., Wei, X., Zhang, S., Duan, H., Cao, M., Zhang, W., Li, Y., Yan, H., Gao, Y., Zhang, X., Li, W., Li, J., Chen, K., He, C., Zhang, X., Qiao, Y., Lin, D., Wang, J.: Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420 (2024)
27. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)

28. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
29. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data (2023)
30. Fu, X., He, M., Lu, Y., Wang, W.Y., Roth, D.: Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? arXiv preprint arXiv:2406.07546 (2024)
31. Fu, X., Zhang, S., Kwon, G., Perera, P., Zhu, H., Zhang, Y., Li, A.H., Wang, W.Y., Wang, Z., Castelli, V., Ng, P., Roth, D., Xiang, B.: Generate then select: Open-ended visual question answering guided by world knowledge. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 2333–2346. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.147, https://aclanthology.org/2023.findings-acl.147
32. Fu, X., Zhou, B., Chandratreya, I., Vondrick, C., Roth, D.: There's a time and place for reasoning beyond the image. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1138–1149. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.81, https://aclanthology.org/2022.acl-long.81
33. Fu, X., Zhou, B., Chen, S., Yatskar, M., Roth, D.: Interpretable by design visual question answering. arXiv preprint arXiv:2305.14882 (2023)
34. 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: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
35. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models (2023)
36. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019)
37. Harris, C., Stephens, M., et al.: A combined corner and edge detector. In: Alvey vision conference. vol. 15, pp. 10–5244. Citeseer (1988)
38. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
39. Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N.A., Luo, J.: Promptcap: Prompt-guided task-aware image captioning. arXiv preprint arXiv:2211.09699 (2022)
40. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. arXiv preprint arXiv:2303.11897 (2023)
41. Hu, Y., Stretcu, O., Lu, C.T., Viswanathan, K., Hata, K., Luo, E., Krishna, R., Fuxman, A.: Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. arXiv preprint arXiv:2312.03052 (2023)
42. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)

43. Lai, Z., Purushwalkam, S., Gupta, A.: The functional correspondence problem. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15772–15781 (2021)
44. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench-2: Benchmarking multimodal large language models. arXiv preprint arXiv:2311.17092 (2023)
45. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
46. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
47. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
48. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
49. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
50. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
51. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
52. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
53. 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? (2023)
54. Liu, Y., Li, Z., Li, H., Yu, W., Huang, M., Peng, D., Liu, M., Chen, M., Li, C., Jin, L., et al.: On the hidden mystery of ocr in large multimodal models. arXiv preprint arXiv:2305.07895 (2023)
55. Liu, Y., Li, Z., Yang, B., Li, C., Yin, X., lin Liu, C., Jin, L., Bai, X.: On the hidden mystery of ocr in large multimodal models (2024)
56. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1150–1157. Ieee (1999)
57. Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A.: Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. arXiv preprint arXiv:2312.17172 (2023)
58. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
59. Marr, D.: Vision: A computational investigation into the human representation and processing of visual information. MIT press (2010)
60. Marr, D., Poggio, T.: Cooperative computation of stereo disparity: A cooperative algorithm is derived for extracting disparity information from stereo image pairs. Science **194**(4262), 283–287 (1976)
61. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv prepreint arXiv:1908.10543 (2019)

62. Minsky, M., Papert, S.: An introduction to computational geometry. Cambridge tiass., HIT **479**(480), 104 (1969)
63. OpenAI: Gpt-4 technical report (2023)
64. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
65. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
66. Sarkar, A., Mai, H., Mahapatra, A., Lazebnik, S., Bhattad, A.: Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28140–28149 (2024)
67. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
68. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European Conference on Computer Vision. pp. 146–162. Springer (2022)
69. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. arXiv preprint arXiv:2304.06712 (2023)
70. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. CVPR (2021)
71. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
72. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. arXiv preprint arXiv:2306.03881 (2023)
73. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
74. Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM (2023)
75. Team, M.N.: Introducing mpt-7b: A new standard for open-source, commercially usable llms (2023), www.mosaicml.com/blog/mpt-7b, accessed: 2023-05-05
76. Torralba, A., Oliva, A.: Depth estimation from image structure. IEEE Transactions on pattern analysis and machine intelligence **24**(9), 1226–1238 (2002)
77. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
78. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., et al.: Muirbench: A comprehensive benchmark for robust multi-image understanding. arXiv preprint arXiv:2406.09411 (2024)
79. Wang, J.Y., Adelson, E.H.: Layered representation for motion analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 361–366. IEEE (1993)
80. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models (2023)

81. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
82. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. arXiv preprint arXiv:2303.09295 (2023)
83. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2022)
84. Yan, A., Yang, Z., Wu, J., Zhu, W., Yang, J., Li, L., Lin, K., Wang, J., McAuley, J., Gao, J., et al.: List items one by one: A new data source and learning paradigm for multimodal llms. arXiv preprint arXiv:2404.16375 (2024)
85. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
86. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
87. Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of gpt-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3081–3089 (2022)
88. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 **9** (2023)
89. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023)
90. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502 (2023)
91. Ze, Y., Wang, X.: Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. Advances in Neural Information Processing Systems **35**, 27469–27483 (2022)
92. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
93. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems **36** (2024)