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]†, 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.[1]

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.

*Both authors contributed equally to this work. Correspondance to <Xingyu Fu: xingyuf2@seas.upenn.edu>. All data and evaluation are available on the project page.

†Both authors advised equally.

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" [7, 24, 26]. In this pursuit, early research developed a series of intermediate tasks: they focused on understanding optical properties like reflectance [3, 30], 3D primitives through multi-view reasoning [13, 25], geometric reasoning through depth estimation [29], instance recognition through visual correspondence [21], affordance through keypoint grounding [12], and forensics through intrinsic images [1]. 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 [4, 6, 8, 22, 27, 28]. 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 [17, 20, 23, 32]. 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 [20]: "`<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 [2, 14, 31]. 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. As partially demonstrated in Figure 1,[1] 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 [9, 10, 15, 19], encompassing indoor household scenes as well as outdoor urban or natural environments. The questions and choices are either derived from the datasets, or 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.

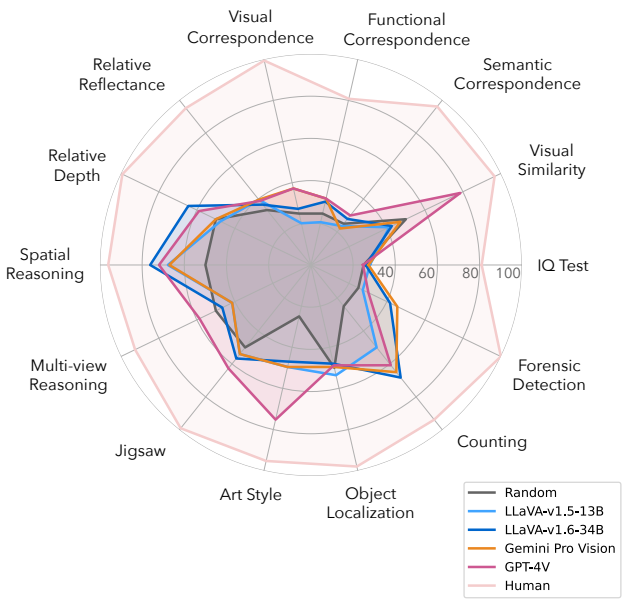

Figure 2. Accuracies of multimodal LLMs on BLINK test set. Please refer to project page for more results and discussions.

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

[1]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.

Figure 3. **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 [32] require domain knowledge. The samples of previous benchmarks are from [18, 20, 32]. Part of our samples are curated from [5, 10, 11, 16, 33].

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. Please refer to the project page for more experiment results.

References

[1] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978. 2

[2] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*, 2023. 2

[3] Michael J Black and Padmanabhan Anandan. A framework for the robust estimation of optical flow. In *1993 (4th) International Conference on Computer Vision*, pages 231–236. IEEE, 1993. 2

[4] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2

[5] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 3

[6] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali-x: On scaling up a multilingual vision and language model, 2023. 2

[7] PHILOSO EPHY DO CT OR OF. *MACHINE PERCEPTION OF THREE-DIMENSIONAL, SO LIDS*. PhD thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 1961. 2

[8] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024. 2

[9] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data, 2023. 2

[10] Xingyu Fu, Ben Zhou, Ishaan Chandratreya, Carl Vondrick, and Dan Roth. There's a time and place for reasoning beyond the image. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1138–1149, Dublin, Ireland, 2022. Association for Computational Linguistics. 2, 3

[11] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 3

[12] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, pages 10–5244. Citeseer, 1988. 2

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[14] Yushi* Hu, Hang* Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022. 2

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2

[16] Zihang Lai, Senthil Purushwalkam, and Abhinav Gupta. The functional correspondence problem. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15772–15781, 2021. 3

[17] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023. 2

[18] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 3

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

[20] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2023. 2, 3

[21] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1150–1157. Ieee, 1999. 2

[22] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023. 2

[23] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2

[24] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 2

[25] David Marr and Tomaso Poggio. Cooperative computation of stereo disparity: A cooperative algorithm is derived for extracting disparity information from stereo image pairs. *Science*, 194(4262):283–287, 1976. 2

[26] Marvin Minsky and Seymour Papert. An introduction to computational geometry. *Cambridge tiass., HIT*, 479(480): 104, 1969. 2

[27] OpenAI. Gpt-4 technical report, 2023. 2

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

[29] Antonio Torralba and Aude Oliva. Depth estimation from image structure. *IEEE Transactions on pattern analysis and machine intelligence*, 24(9):1226–1238, 2002. 2

[30] John YA Wang and Edward H Adelson. Layered representation for motion analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 361–366. IEEE, 1993. 2

[31] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3081–3089, 2022. 2

[32] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 2, 3

[33] Yanjie Ze and Xiaolong Wang. 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. 3