Visual Cropping Improves Zero-Shot Question Answering of Multimodal Large Language Models

Jiarui Zhang
Information Sciences Institute
University of Southern California, USA
jrzhang@isi.edu

Mahyar Khayatkhoei
Information Sciences Institute
University of Southern California, USA
mkhayat@isi.edu

Prateek Chhikara
Information Sciences Institute
University of Southern California, USA
pchhikar@isi.edu

Filip Ilievski
Department of Computer Science
Vrije Universiteit Amsterdam, Netherlands
f.ilievski@vu.nl

Abstract

Multimodal Large Language Models (LLMs) have recently achieved promising zero-shot accuracy on visual question answering (VQA) – a fundamental task affecting various downstream applications and domains. Given the great potential for the broad use of these models, it is important to investigate their limitations in dealing with different image and question properties. In this work, we investigate whether multimodal LLMs can perceive small details as well as large details in images. In particular, we show that their zero-shot accuracy in answering visual questions is very sensitive to the size of the visual subject of the question, declining up to 46% with size. Furthermore, we show that this effect is causal by observing that human visual cropping can significantly mitigate their sensitivity to size. Inspired by the usefulness of human cropping, we then propose three automatic visual cropping methods as inference time mechanisms to improve the zero-shot performance of multimodal LLMs. We study their effectiveness on four popular VQA datasets, and a subset of the VQAv2 dataset tailored towards fine visual details. Our findings suggest that multimodal LLMs should be used with caution in detail-sensitive VQA applications, and that visual cropping is a promising direction to improve their zero-shot performance. Our code and data are publicly available. [1]

1 Introduction

Visual question answering (VQA) is a fundamental task with a broad range of downstream applications in many critical domains, from biomedicine [19, 14, 9] to traffic monitoring [23, 25] and remote sensing [17, 13]. Zero-shot VQA – answering visual questions in a domain without having access to annotated data from that specific task and domain – is of particular interest since collecting reliable answers for an extensive number of question-image pairs is expensive and time-consuming, and thus impractical for many downstream tasks due to lack of access to experts or privacy and security concerns [26]. Recently, multimodal Large Language Models (LLMs) [11, 1, 15] have shown promising accuracy in zero-shot VQA, commonly attributed to their pretraining on terabytes of image and language data with billion-parameter Transformer-based neural networks. Given their potentially broad adoption in downstream tasks, it is crucial to study their limitations in dealing with various phenomena in images and questions. To that end, in this work, we investigate whether their question-answering ability is affected by the size of the visual object of interest.

[1] https://github.com/saccharomycetes/visual_crop_zsvqa

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

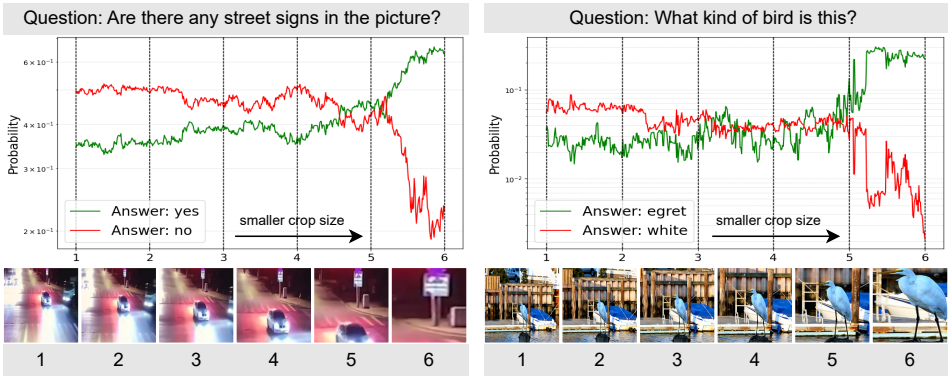


Figure 1: The effect of visual cropping on the probability of answers predicted by BLIP2 in zero-shot VQA. The x-axis represents the crop size around the relevant visual subject of the question (x-axis labels are indices to the cropped images displayed below each plot that the model sees at each step). The model gradually finds the correct answer as it looks closer and perceives smaller visual details.

In Figure 1, we provide two motivating examples to illustrate a limitation in multimodal LLMs that we will study in this paper in more detail. In these examples, we ask BLIP2 (FlanT5XL) [11], the state-of-the-art multimodal LLM in zero-shot VQA, two questions about relatively small objects in the image, *i.e.*, questions concerning *visual details*. In the absence of any prior empirical evidence, it is reasonable to assume that the BLIP2 accuracy is not significantly affected by the size of the question's visual subject because of the large representational capacity of multimodal LLMs and their pretraining on a large variety of images containing objects of various sizes. In Figure 1 (left), we observe that initially the model does not recognize the existence of a small street sign and assigns lower probability to the correct answer; however, zooming into the image towards the street sign gradually increases the probability assigned to the correct answer, suggesting that the model gradually perceives more and more relevant details of the street sign. In Figure 1 (right), we observe further evidence of this limitation in perceiving visual details. The model initially predicts *white* as the type of the bird; however, when we zoom into the image towards the bird via visual cropping, without changing the question in any way, we observe that the model gradually assigns higher probability to the correct bird type of *egret*, suggesting that the model was not making a semantic error of misunderstanding what *type* means, neither was it confused about where to look, rather it was unable to perceive sufficient details to discriminate egret from other white birds, which is enabled by visual cropping. This observation is particularly surprising since the visual encoding in BLIP2 is not theoretically restricted in its visual resolution and therefore should be able to perceive the traffic sign and recognize the bird type regardless of their relative visual sizes. The main goal of this paper is to investigate the extent of the limitation observed in Figure 1 and potential solutions to mitigate its consequences. Our study will focus on two variants of BLIP2 as the state-of-the-art for zero-shot VQA (outperforming Flamingo [1]) and the only fully open-sourced multimodal LLM.

2 Related Works

Multimodal LLMs can be broadly grouped into two categories: end-to-end pretrained models, and modular pretrained models. The former group includes architectures that are explicitly designed for processing joint image and language data, most notably, the dual-encoder [16], the fusion-encoder [10], the encoder-decoder [3], and the unified transformer [22], which are trained with common pretraining objectives: image-text matching and contrastive, and masked language modeling. The second group aims to overcome the expensive pretraining cost of the former group by learning to adapt existing pretrained models: some models use a frozen image encoder and finetune an LLM with the pretraining objectives [24, 27], whereas some models instead freeze the LLM and finetune the vision encoder with additional adaptor layers [1, 21]. The most successful such model is BLIP2 [11], which freezes both the vision encoder and the LLM, and directly learns a transformer-based module on pretraining objectives to bridge the modality gap of its frozen underlying models. The granular sensitivity of multimodal LLMs, including BLIP2, to image and question properties, has not been thoroughly studied. We aim to bridge this gap in the present paper.

Table 1: Sensitivity of zero-shot accuracy of VQA models to the size of visual concepts in TextVQA. As the relative visual size of the answer decreases (right to left in each row), we observe a significant decline in the accuracy of original models, whereas visual cropping reduces this accuracy gap.

Model	Crop Method	Answer Bounding Box Size (S)		
		< 0.005	$[0.005, 0.05)$	≥ 0.05
BLIP2 (FlanT5$_{XL}$)	w/o cropping	19.91	29.07	36.81
	`human-CROP`	32.06	41.31	38.84
BLIP2 (OPT$_{2.3B}$)	w/o cropping	19.38	26.09	33.28
	`human-CROP`	27.19	34.36	33.25

3 Sensitivity of Zero-Shot VQA Models to the Size of Visual Concepts

In this section, our goal is to quantitatively test our qualitative observations in Figure 1 that the zero-shot VQA models struggle with describing fine visual details in images. To that end, we consider the Text-VQA dataset, where for each question we can find the ground-truth bounding box containing the correct answer (detailed in Appendix B). We partition its validation set into three groups based on the relative size of the ground truth bounding box $S = \frac{A_{bb}}{A_{total}}$, where A_{bb} denotes the area of the answer bounding box, and A_{total} denotes the total area of the image: 1) $S < 0.005$ consisting of 2822 question-image pairs, 2) $0.005 \leq S < 0.05$ consisting of 1833 question-image pairs, and 3) $S \geq 0.05$ consisting of 345 question-image pairs. If a model's perception is not sensitive to the size of visual concepts, we expect it to have similar accuracy in all three groups. In Table 1, we observe that the accuracy of both BLIP-2 variants declines across the three groups as the answer bounding box becomes smaller (see `w/o cropping` rows). BLIP-2 (FlanT5) exhibits an accuracy decline of 46% from group 3 (largest visual concepts) to group 1 (smallest visual concepts), and BLIP-2 (OPT) exhibits a similar decline of 42%. These findings show that both models answer questions about visual concepts more accurately when their relative size is larger, *i.e.*, they struggle to perceive fine visual details. Furthermore, to confirm that the issue is causally related to the size of the visual concept, we conduct an intervention study, where we provide the models with visually cropped images based on the ground-truth bounding boxes, denoted as `human-CROP`. We observe in Table 1 that `human-CROP` significantly improves the accuracy of both models, but more importantly, under `human-CROP` the accuracy across the three groups becomes more similar: the decline between the largest visual concept setting (group 3) to the smallest one (group 1) is less than 19% for both models. This suggests that the perception limitation is indeed caused by the size of the visual concepts, and that visual cropping can improve the perception of visual details by the zero-shot VQA models.

4 Visual Cropping Methods

To investigate whether the accuracy gain achieved by human visual cropping in Section 3 is realizable in practice, in this section we propose three automatic cropping methods, illustrated in Figure 2, whose goal is to find the approximate region of interest in images, *i.e.* the region containing the subject of a question, and then to zoom into that region via visual cropping: 1) `clip-CROP` where we progressively crop the image towards the region of highest relevance to a given question using CLIP [16]; 2) `yolo-CROP` where instead of progressive cropping, we use candidate bounding boxes predicted by YOLOv8 [6] to filter out regions that contain no salient objects, *i.e.*, regions for which CLIP could mistakenly assign high similarity; 3) `sam-CROP` where instead of YOLO, which provides bounding boxes for only a fixed number of object classes, we use the segment anything model (SAM) [7] to provide an extensive set of salient segmentation masks for each image as candidate regions to compare to the question using CLIP. One potential drawback of visual cropping is that some questions might need to have a global view of the image. To address this issue, we utilize the fact that multimodal LLMs typically convert the image into a series of tokens. This allows us to directly extend the original image tokens by concatenating the visually cropped image tokens, as illustrated in Figure 2. More details about the methods, and examples of their success and failures are provided in Appendix. In the next section, we will investigate the effectiveness of the proposed methods for improving the accuracy of zero-shot VQA models.

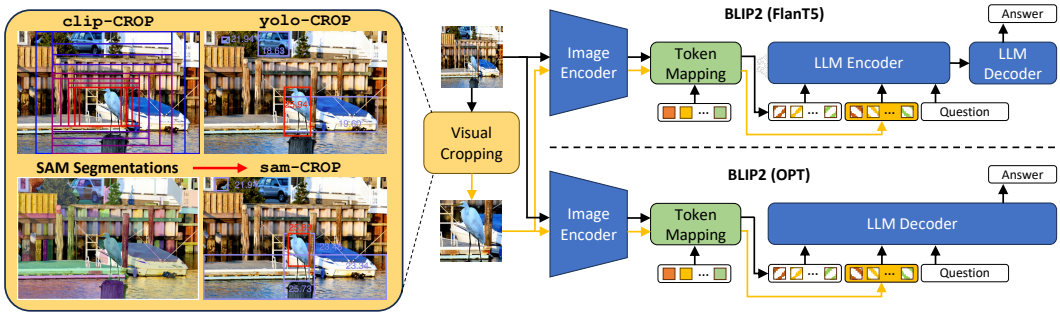


Figure 2: Illustration of the proposed visual cropping methods applied to two variants of BLIP2.

Table 2: Accuracy of human and automatic visual cropping methods on VQA datasets. For each dataset and model, the best cropping method is depicted in bold, and the second-best is underlined.

Model	Crop Method	FD-VQA	TextVQA	GQA	A-OKVQA	VQAv2
BLIP2 (FlanT5$_{XL}$)	w/o cropping	33.94	25.91	43.85	<u>43.42</u>	63.43
	`clip-CROP`	<u>36.61</u>	30.93	45.04	**45.34**	<u>63.57</u>
	`yolo-CROP`	35.87	28.94	<u>45.20</u>	42.72	63.39
	`sam-CROP`	36.33	<u>32.31</u>	**45.23**	43.00	**63.85**
	`human-CROP`	**42.29**	**37.68**	-	-	-
BLIP2 (OPT$_{2.3B}$)	w/o cropping	35.14	23.93	**31.95**	<u>31.57</u>	**51.22**
	`clip-CROP`	<u>35.60</u>	26.35	<u>31.14</u>	**32.69**	<u>49.67</u>
	`yolo-CROP`	34.86	25.27	30.39	29.43	48.63
	`sam-CROP`	35.60	<u>26.45</u>	30.47	31.07	48.58
	`human-CROP`	**42.11**	**31.21**	-	-	-

5 Experiments

In this section, we investigate the effectiveness of the proposed visual cropping methods in improving the zero-shot accuracy of VQA models. We use two state-of-the-art zero-shot VQA models which have open-source code [11]: *BLIP2 (FlanT5XL)* is an encoder-decoder LLM and, *BLIP2 (OPT2.3B)*, a decoder-only LLM. We experiment with four popular VQA datasets, VQAv2 [4], GQA [5], A-OKVQA [18], TextVQA [20], and a new fine-detail subset of VQAv2, denoted FDVQA, to enrich our evaluation of questions about visual details. Details of the datasets are provided in Appendix B.

Table 2 shows the accuracy[2] of the proposed visual cropping methods on the five VQA datasets. First, we consider the detail-focused datasets, FDVQA and TextVQA, where we also have access to human annotations and can report `human-CROP` accuracy: we observe that `human-CROP` improves the accuracy of both the BLIP2 FlanT5 and OPT models, by 24% and 20%, respectively, showing the full potential of visual cropping; additionally, for the proposed visual cropping methods, we observe that while they do not achieve the full potential of `human-CROP`, they still successfully improve upon both original models, with `clip-CROP` achieving best improvement on FDVQA and `sam-CROP` on TextVQA. Next, we consider GQA, A-OKVQA, and VQAv2. For BLIP2 (FlanT5), we observe that visual cropping methods can improve the accuracy of the original models, which shows that their accuracy gain on fine details (observed in FDVQA and TextVQA) does not come at the cost of their accuracy on larger visual details and relations. However, for BLIP2 (OPT), we observe that visual cropping methods can cause a decline in the overall accuracy. We hypothesize that this is due to the use of a decoder-only Transformer architecture in OPT which expects visual tokens to appear at specific initial positions and therefore could not correctly attend to a concatenation of two sets of image tokens (see Figure 2). We provide additional results regarding the inference time and sensitivity of the methods to various question types in Appendix. Overall, our findings suggest that: 1) in settings where fine visual details are the main subject of the questions, visual cropping is a promising technique to improve zero-shot VQA accuracy; and 2) there is a need for better visual cropping methods that can close the gap with human visual cropping.

[2]https://visualqa.org/evaluation.html

Acknowledgments and Disclosure of Funding

This work has been supported by the Keston ISI Exploratory Research award "Coherent and commonsensical AI for social influence" and by DARPA's Machine Common Sense (MCS) program.

References

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

[2] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 71–79, 2018.

[3] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

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

[6] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023. URL `https://github.com/ultralytics/ultralytics`.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[8] 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.

[9] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[10] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[12] 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.

[13] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

[14] Usman Naseem, Matloob Khushi, and Jinman Kim. Vision-language transformer for interpretable pathology visual question answering. *IEEE Journal of Biomedical and Health Informatics*, 27(4):1681–1690, 2022.

[15] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

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

[17] Argho Sarkar and Maryam Rahnemoonfar. Vqa-aid: Visual question answering for post-disaster damage assessment and analysis. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 8660–8663. IEEE, 2021.

[18] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer, 2022.

[19] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 33–43. Springer, 2022.

[20] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[21] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212, 2021.

[22] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[23] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9878–9888, 2021.

[24] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[25] Jiarui Zhang, Filip Ilievski, Kaixin Ma, Aravinda Kollaa, Jonathan Francis, and Alessandro Oltramari. A study of situational reasoning for traffic understanding. *KDD*, 2023.

[26] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. How to practice vqa on a resource-limited target domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4451–4460, 2023.

[27] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

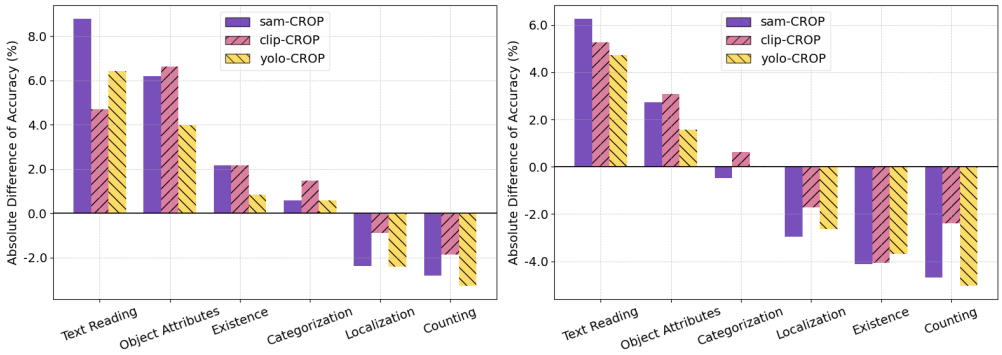


Figure C.1: Accuracy gain of visual cropping methods compared to no cropping, when applied to BLIP2 FlanT5 (left) and OPT (right) on different question types in VQAv2. The x-axis is sorted based on the combined gain of all methods. The question types are explained in Table 3 in Appendix.

A Implementation Details

We use *python 3.8.16, salesforce-lavis 1.0.2, transformers 4.29.1 and torch 2.0.1* for all the BLIP model experiments. Our environment consists of an Intel(R) Xeon(R) Gold 5215 CPU @ 2.50GHz with 40 cores and 256 GB of RAM. Additionally, we utilize NVIDIA RTX A5000 GPUs for our experiments.

B Dataset Details

We consider the validation set of four common VQA datasets and construct a new one tailored towards visual details: 1) **VQAv2** [4], which is a large-scale dataset containing 214,354 questions paired with 40,504 images from various objects and settings (a subset of COCO [12]); 2) **GQA** [5] containing 12578 questions paired with 398 images, constructred by using the scene graphs of Visual Genome [8] to construct highly compositional questions requiring spatial, logical, relational, and comparative reasoning, and explicitly controls the answer distribution for different groups of questions in order to prevent educated guesses using language and world priors; 3) **A-OKVQA** [18], which contains containing 1145 questions about 1122 images, where the questions require additional knowledge and cannot be answered from the image-question pair alone; 4) **TextVQA** [20] containing 5000 questions about textual information that appear in 3166 images, where more than half of the answers require perceiving texts that occupy less than 0.005 of the total image area, and therefore emphasizes how well a model can read small text, which can serve as a surrogate for how well a model can perceive fine visual details; additionally, textVQA provides Optical Character Recognition (OCR) annotations [2] which we use to approximate the ground-truth answer bounding box for each question by selecting the OCR bounding box containing the text with the highest string similarity with the human-provided answer. This bounding box is used for cropping in `human-CROP`; 5) **FDVQA** is a dataset we propose to deliberately focus on small hard-to-perceive visual details; for this purpose, we first selected 400 question-answer pairs of VQAv2 on which the zero-shot BLIP2 model fail to correctly predict the majority answer in the annotations, in order to filter out any sample where perception is easy; then, we collected 3 human annotations per sample identifying whether answering the question requires perceiving small details in the image and the model answer is indeed incorrect (*e.g.*, excluding near-synonymous answers or ambiguous questions); finally, we kept the samples where all 3 annotations agreed, resulting in 109 image-question pairs, and we manually created the ground-truth bounding box around the subject of the question.

C The Effect of Visual Cropping On Different Question Types.

To gain deeper insights into the granular benefits of visual cropping, Figure C.1 shows how the proposed visual cropping methods impact the accuracy of zero-shot VQA models on various question types in VQAv2 (these types are explained in Table 3 in Appendix). Questions concerning visual details, *i.e.*, text reading and object attributes, gain the most from visual cropping, consistent with our findings in FDVQA and TextVQA. However, we observe that questions that require a global view of the image, *i.e.*, localization and counting, become harder to answer as a result of visual cropping. This suggests that our mechanism for combining the original and cropped image tokens is not always successful in maintaining the global image information, encouraging future research on effectively combining image tokens.

Table 3: We select 6 question types from VQAv2 based on their first two words to study the granular accuracy of visual cropping methods in Section 5. The total number of instances per question type is reported in the last row, with 140691 instances belonging to none of these types.

Question types					
Reading	**Object Attributes**	**Existence**	**Categorization**	**Localization**	**Counting**
what letter	what pattern	is anyone	what street	where is	how many
what brand	what color	is there	what direction	where are	how much
	what breed	are there	what animal	where was	
	what colors	is that	what fruit		
	what style	are all	what vegetable		
	what material	is everyone	what food		
	what shape	is one	what game		
		is she	what sport		
		is he			
1064	22053	16426	4168	6329	23623

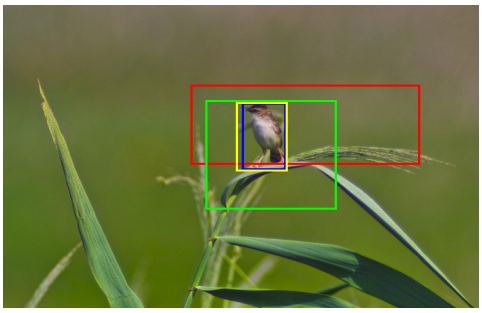
Table 4: Inference time comparison of visual cropping methods.

Cropping Method	`clip-CROP`	`yolo-CROP`	`sam-CROP`	DEVICE
Average Inference Time (seconds)	1.072	0.355	3.329	GPU
	5.461	0.970	91.532	CPU

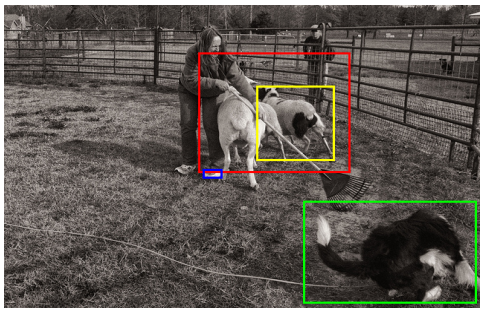
D The Time Overhead of Visual Cropping.

In Table 4, we report the average inference time of the proposed visual cropping methods on GPU (NVIDIA RTX A5000) and CPU (Intel Xeon Gold 5215 2.50GHz). While YOLO is the fastest method, we recommend using CLIP in practice since it provides the best overall balance between accuracy and inference time.

E Additional Examples of Visual Cropping Success and Failure



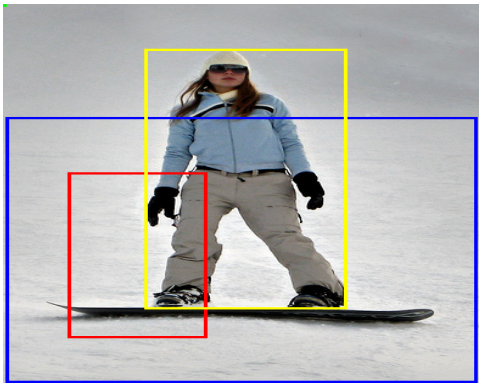
(a) FD-VQA ✓
Question: What is the bird sitting on?
Answer: Leaf



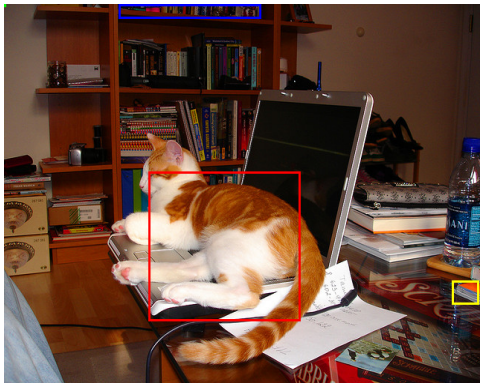
(b) FD-VQA ✗
Question: How many dogs are here?
Answer: 1

(c) TextVQA ✓
Question: What letter does these athelete's school likely begin with?
Answer: G

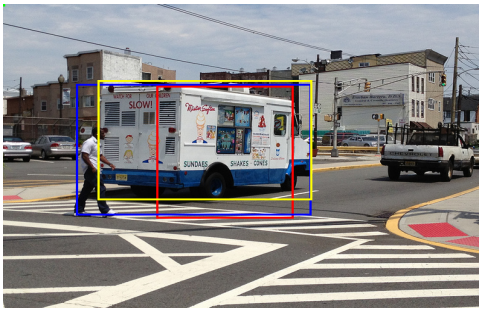
(d) TextVQA ✗
Question: Who is winning?
Answer: Michigan



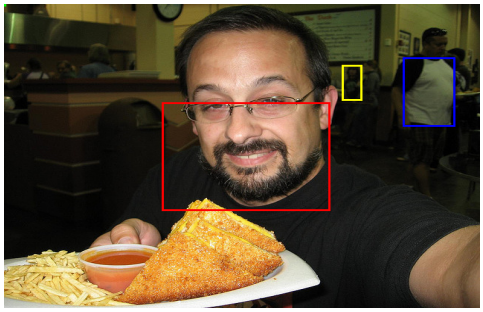
(e) GQA ✓
Question Is the person wearing a glove?
Answer: Yes



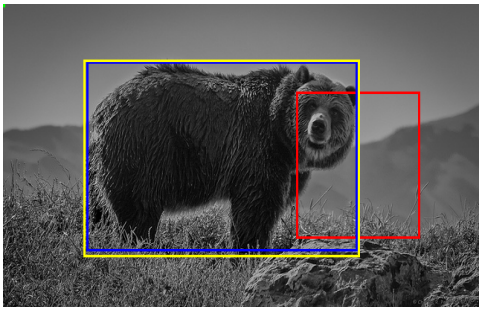
(f) GQA ✗
Question: What item of furniture is large?
Answers: Desk



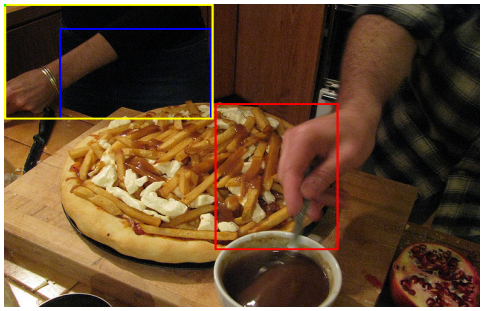
(g) A-OKVQA ✓
Question: What does the ice-cream truck say to watch out for?
Answer: Children



(h) A-OKVQA ✗
Question: What is the white object the man in the black shirt is holding?
Answer: Plate



(i) VQAv2 ✓
Question: Is the mouth open?
Answer: No



(j) VQAv2 ✗
Question: How many people do you see?
Answer: 2

Figure E.2: Success (✓) and Failures (✗) of the proposed cropping techniques on five different datasets.