BRAVE🦁: Broadening the visual encoding of vision-language models
Supplementary Material

Oğuzhan Fatih Kar[1,2] Alessio Tonioni[1] Petra Poklukar[1]
Achin Kulshrestha[1] Amir Zamir[2] Federico Tombari[1]

[1]Google [2]Swiss Federal Institute of Technology Lausanne (EPFL)

https://brave-vlms.epfl.ch

Table of Contents

1 Overview of evaluation tasks 1
2 Training details 1
2.1 Pre-training 1
2.2 Fine-tuning 2
2.3 Prompt examples 3
3 Additional results 3
3.1 Qualitative comparisons 3
3.2 Ablations 3
3.3 Contribution of vision encoders 5

1 Overview of evaluation tasks

Figure 1 provides an overview of the evaluation tasks we used in our paper. They include a broad range of tasks from captioning and visual question answering, and assess different capabilities of the VLMs such as spatial reasoning, robustness to visual hallucinations, novel object captioning, etc. Please see the paper for more details.

2 Training details

2.1 Pre-training

For the main results, we use an image resolution of 224×224, batch size 1024, and 0.1 dropout for the LM during pre-training. We train the VLMs for 900k steps with peak learning rate $5e-5$ with 2k linear warm-up steps, followed by cosine learning rate decay. We use AdamW [15] as the optimizer for all training stages. The pre-training is done using 64 TPUv5 chips and takes around 2 days. During pre-training, both vision encoder(s) and the language model are kept frozen, and only the parameters of the `MEQ-Former` are trained.

Fig. 1: Overview of the evaluation tasks we used in our paper.

2.2 Fine-tuning

Captioning. We perform fine-tuning on COCO training set at 336×336 image resolution by keeping the vision encoder(s) and the LM frozen. We use batch size 64 and train for $20k$ steps with 0.1 dropout on the LM. The peak learning rate is set to $1e-5$ with $2k$ linear warmup, followed by cosine learning rate decay. The resulting model is evaluated both on COCO evaluation sets and on NoCaps (zero-shot), as explained in the main paper.

VQA. For the VQA-mixture, we use samples from VQAv2 [8] and OKVQA [16] as well as the synthetically generated VQA data from VQ^{2}A [3]. For VQ^{2}A, we use the synthetic samples generated on both COCO [13] and CC3M [4] training sets, which amounts to $17M$ examples in total. We use the following mixture ratio: $\{10, 1, 10, 10\}$ to sample from VQAv2, OKVQA, VQ^{2}A-COCO and VQ^{2}A-CC3M, respectively. This training is performed for $60k$ steps with a batch size of 1024 and 0.1 dropout for the LM. Both the `MEQ-Former` and the LM parameters are updated while vision encoders are always kept frozen. The peak learning rate is set to $1e-5$ with $5k$ linear warm-up steps, followed by cosine learning rate decay. This stage is performed at 224×224 image resolution, and is followed by a high-resolution fine-tuning stage at 336×336 resolution on the VQAv2 training set. For this, we train for $20k$ steps with a batch size of 1024. The peak learning rate is set to $1e-5$ with $4k$ linear warm-up, followed by cosine learning rate decay. For fine-tuning on GQA [9] training set, we follow the same recipe by only changing the total number of training steps to $10k$ and warm-up to $2k$ steps. The resulting models are evaluated on several VQA benchmarks (both zero-shot and fine-tuned cases), as explained in the main paper.

Table 1: Example prompts used for evaluations on different tasks and datasets. BRAVE can make use of different types of prompts, even those that were not seen during training, showing zero-shot prompt generalization capabilities.

Dataset	Prompt
COCO, NoCaps	A photo of
VQAv2, OKVQA, GQA	<Question>
VizWiz-QA	Answer as "Unanswerable" when the image content is not clear. Question: <Question>
MMVP	<Question>. Answer <option 1> or <option 2>.
POPE	Answer as yes or no. Question: <Question>

2.3 Prompt examples

We use different prompts for BRAVE based on the task and dataset, as summarized in Table 1. Note that some of the prompts were not seen during training, e.g. VizWiz-QA or MMVP, yet the model is able to generalize to them zero-shot, showing that we preserve the language understanding capabilities of the LLM, as also demonstrated in [2, 14].

3 Additional results

3.1 Qualitative comparisons

Captioning. We perform zero-shot captioning on NoCaps [1] validation set images in Figure 2. The results show that BRAVE creates accurate descriptions for a diverse set of inputs with visual abstractions, novel classes, and fine-grained details. The quantitative evaluations in Table 3 of the main paper further confirms the improvements achieved by BRAVE.

Visual question answering. We show in Fig. 3 additional comparisons of BRAVE against InstructBLIP [7] and LLaVA-1.5 [14] as well as single vision encoder VLMs (from Sec. 2 of the main paper) on MMVP [17] test pairs, further demonstrating the improved performance for a diverse set of inputs. We also observe that some inputs remain challenging for all VLMs, e.g. those require fine-grained text or scene understanding, which can benefit from incorporating additional biases and is a possible future work direction.

3.2 Ablations

MEQ-Former vs Ensembling. In the main paper, we discussed that resampling all vision encoders by MEQ-Former lead to strong performance while being efficient in terms of the number of trainable parameters. In Table 2, we compare MEQ-Former to an ensemble of Q-Formers [11]. Each Q-Former is first fully pre-trained with its corresponding single vision encoder. This is followed by a joint pre-training and a fine-tuning stage by ensembling all vision encoders and their pre-trained Q-Formers. The outputs of all Q-Formers are fed as input to the same

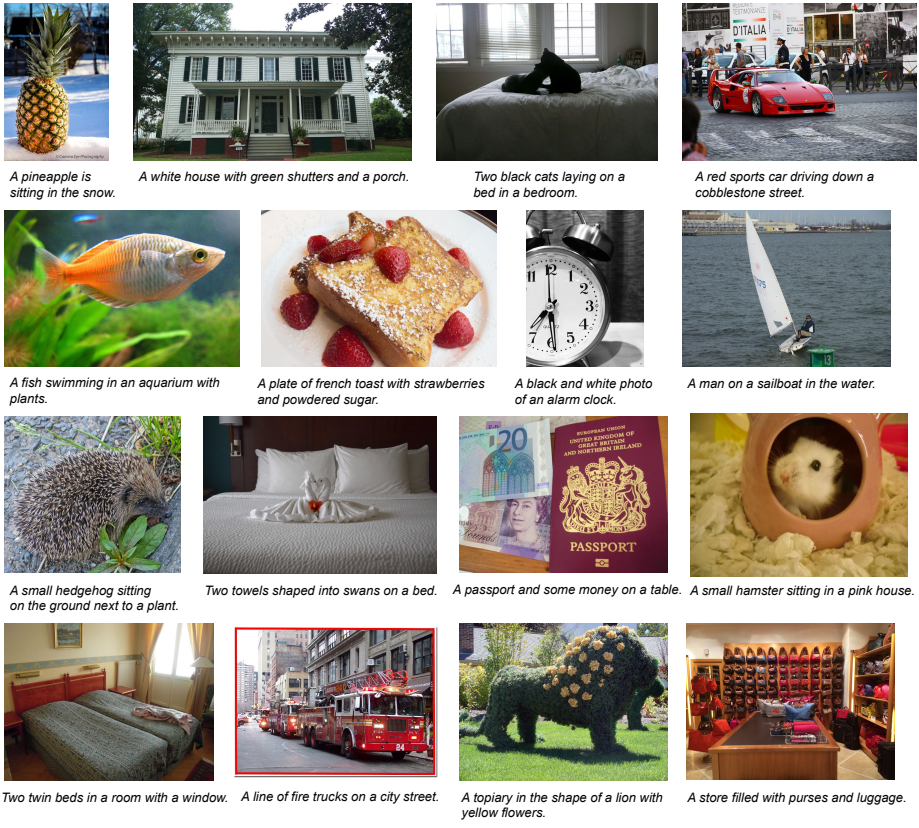

Fig. 2: Qualitative results on captioning. We perform zero-shot captioning on samples from NoCaps [1] validation set using `BRAVE`. See Table 3 in the main paper for quantitative evaluations.

LM (FlanT5-XL [6]). We use the same five encoders as in the main paper, hence the only difference is using an ensemble of Q-Formers instead of `MEQ-Former` for resampling, resulting in 5x the number of trainable parameters. Our results in Table 2 show that `MEQ-Former` resamples visual features more effectively for several tasks while using less trainable parameters.

Role of pre-training data. In Table 3 we evaluate the impact of pre-training data by comparing the `BRAVE` models pre-trained on WebLI [5] and CC3M [4] datasets. The latter has about 30× less samples than the former, and pre-training with it leads to a noticeable degradation in performance, suggesting that more work is needed to reduce the sample complexity of VLMs, e.g. as studied in [10].

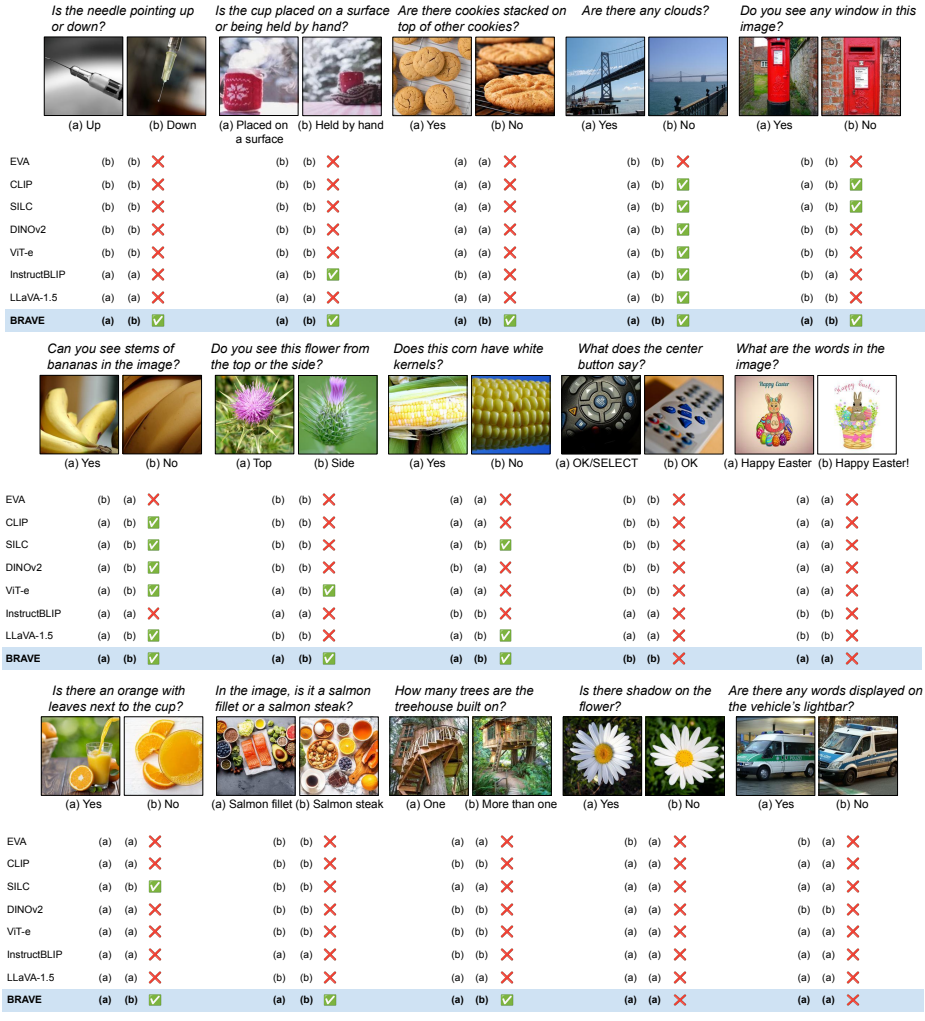

Fig. 3: Qualitative results on VQA. This is an extension of Fig. 3 in the main paper. Example pairs are taken from [17]. `BRAVE` significantly improves performance for a broad set of challenging inputs compared to recent methods [7, 14] as well as single vision encoder based VLM baselines. The improvement can also be seen quantitatively in Table 4 in the main paper. On the other hand, some examples remain challenging for all VLMs, e.g. those require fine-grained text or scene understanding, which can benefit from incorporating additional biases targeting them in a future study.

3.3 Contribution of vision encoders

Cross-attention scores. We provide additional cross-attention score visualizations in Figure 4 on NoCaps [1] captioning and POPE [12] visual question answering benchmarks. Similar to Fig. 4 in the main paper, they demonstrate

Table 2: `MEQ-Former` vs Ensembling. We compare resampling vision encoder features by using an ensemble of Q-Formers and `MEQ-Former`. The latter uses significantly less trainable parameters and better captures the strengths of different vision encoders, leading to consistently better performance. All evaluations are performed at 224×224 resolution. See Sec. 3.2 for details.

Bridge	# of parameters	COCO Cap.	VQAv2	OKVQA	GQA
Q-Former Ensemble	605M	140.9	78.5	64.3	50.6
`MEQ-Former`	**116M**	**145.2**	**79.6**	**65.0**	**51.5**

Table 3: Role of pre-training data. We compare VLMs pre-trained on WebLI [5] and CC3M [4] datasets. All evaluations are performed at 224×224 resolution. The latter has significantly less image-text pairs which leads to a degradation in the performance, suggesting more studies are needed to reduce the sample complexity of VLM training. See Sec. 3.2 for details.

Pre-training Dataset	COCO Cap.	VQAv2	OKVQA	GQA
CC3M	138.3	76.9	63.4	50.0
WebLI	**145.2**	**79.6**	**65.0**	**51.5**

that the `MEQ-Former` cross-attends vision encoder features adaptively depending on the downstream task.

Robustness analysis. We provide in Table 4 the full evaluation results of the robustness of `BRAVE` against missing encoders for COCO captioning and VQAv2 visual question answering tasks. It can be seen that the performance degrades gracefully up to 2 encoder removals, and removal of some encoders hurts the performance more than the others, e.g. ViT-e and SILC-G/16, suggesting that their features are harder to replace by the remaining set.

Table 4: Robustness analysis of `BRAVE` against missing encoders. This is an extension of Fig. 4 (left) of the main paper, and shows the performance on COCO captioning and VQAv2 for all combinations of vision encoders. **E:** EVA-CLIP-g **C:** CLIP-L/14 **D:** DINOv2-L/14 **V:** ViT-e **S:** SILC-G/16

COCO Captioning		VQAv2	
Combination	CIDEr	Combination	Accuracy
ECDVS	146.5	ECDVS	81.8
ECVS	146.6	ECVS	81.7
CDVS	146.4	EDVS	81.3
EDVS	146.2	ECDS	81.2
ECDS	145.3	CDVS	81.2
ECDV	142.5	ECDV	80.1
CVS	146.3	EVS	81.4
EVS	146.2	CVS	81.2
EDS	145.1	EDS	81.0
ECS	145.0	DVS	80.9
DVS	145.0	ECS	80.7
CDS	144.8	CDS	80.5
EDV	142.6	ECV	79.7
ECV	141.5	EDV	79.6
CDV	140.4	CDV	78.5
ECD	136.3	ECD	78.3
VS	145.2	CS	80.6
CS	145.1	VS	80.4
ES	144.5	ES	80.2
DS	142.8	DS	79.6
EV	141.6	EV	79.0
CV	139.7	EC	78.2
EC	135.3	CV	78.0
ED	133.5	ED	75.9
DV	131.3	CD	75.6
CD	128.6	DV	73.4
S	141.8	S	79.8
E	129.9	C	75.4
V	128.9	E	75.3
C	126.2	V	72.0
D	45.0	D	53.2

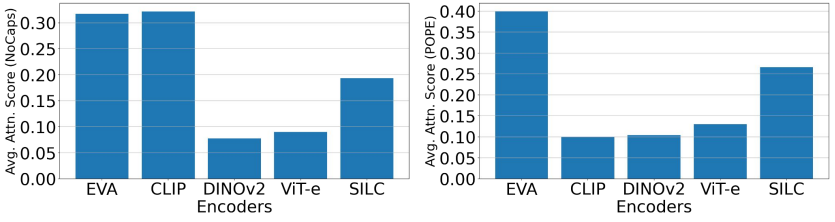

Fig. 4: Contribution of vision encoders to BRAVE. This is an extension of Fig. 4 (right) in the main paper. We compute average attention scores for different vision encoders cross-attended by the `MEQ-Former` for NoCaps (left) and POPE (right). The `MEQ-Former` cross-attends different vision encoder features adaptively, depending on the downstream task.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–8957 (2019) 3, 4, 5
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 3
3. Changpinyo, S., Kukliansky, D., Szpektor, I., Chen, X., Ding, N., Soricut, R.: All you may need for vqa are image captions. arXiv preprint arXiv:2205.01883 (2022) 2
4. Changpinyo, S., Sharma, P.K., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3557–3567 (2021) 2, 4, 6
5. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022) 4, 6
6. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 4
7. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=vvoWPYqZJA 3, 5
8. 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: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017) 2
9. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019) 2

10. Jian, Y., Gao, C., Vosoughi, S.: Bootstrapping vision-language learning with decoupled language pre-training. Advances in Neural Information Processing Systems **36** (2024) 4
11. 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) 3
12. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023) 5
13. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (2014) 2
14. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023) 3, 5
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 1
16. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019) 2
17. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. arXiv preprint arXiv:2401.06209 (2024) 3, 5