Supplementary Materials for FlexAttention for Efficient High-Resolution Vision-Language Models

Junyan Li[1], Delin Chen[1], Tianle Cai[2], Peihao Chen[3], Yining Hong[4], Zhenfang Chen[5], Yikang Shen[5], and Chuang Gan[1,5]

[1] UMass Amherst
[2] Princeton University
[3] South China University of Technology
[4] University of California, Los Angeles
[5] MIT-IBM Watson AI Lab
{junyanli,delinchen}@umass.edu
tianle.cai@princeton.edu
{phchencs,yninghong,chenzhenfang2013,yikang.shn,ganchuang1990}@gmail.com

A Appendix

A.1 Evaluation Datasets

In Sec. 5.3, We conduct experiments on these four high-resolution benchmarks: V* Bench [10], MagnifierBench [6], TextVQA [9] and RSVQA-HRBEN [8].

- **V* Bench** contains 191 high-resolution images from SA-1B dataset [4] with an average resolution of 2246×1582. There are two sub-tasks in this benchmark: attribute recognition and spatial relationship reasoning. Following [10] we select the choice with the lowest perplexity as the model's prediction and compute the accuracy.
- **MagnifierBench** contains 283 QA pairs that have a typical image resolution of 1920x1080. The questions vary widely, covering identification, numerical, color-related inquiries, and more. All questions are multiple-choice, and the model is required to directly answer the option letter from the given choices. Accuracy is computed as the evaluation metric.
- **TextVQA** contains 5000 questions for 3166 images. The average resolution of the images in the dataset is 950x811. It requires the model to read and reason about the text in images to answer the question. We evaluate models in an OCR-free manner, which means that no OCR results are provided. This prevents the model from guessing the answer from the OCR result if it cannot recognize the text correctly. Accuracy is computed following [7].
- **RSVQA-HRBEN** is the high-resolution subset of the RSVQA dataset. It contains 47k question-answer pairs for remote sensing images. The questions are categorized into three types: presence, comparison, and count. Questions in presence category ask whether an object is presented in the image. Questions in comparison category ask to compare the number of object a to the

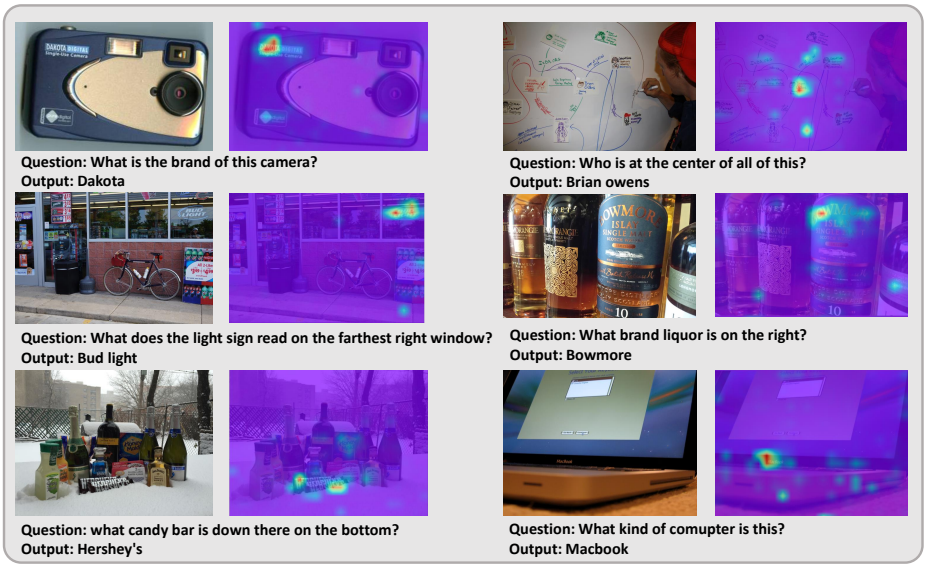

Fig. 1: Visualization of the attention map. The highlighted area indicates the region that the model is paying attention to in the image when answering the question.

number of object b. Questions in count category ask to count the number of a object. Following [5], we evaluate the model on the presence and comparison categories. We report the accuracy of these two categories as well as the average accuracy.

A.2 Attention Map Visualization

We offer a qualitative visualization of the attention map in Fig. 1 to demonstrate the VLM's attention regions while generating tokens.

A.3 Attention Map Quantitative Analysis

In our method, accurately identifying important regions of the high-resolution features via the attention map is crucial. The effectiveness of our method hinges on how precisely the attention map selects regions that enhance question answering. To assess the accuracy of the attention map for revealing important regions, we conduct a quantitative experiment on GQA [3] dataset using the identical evaluation script provided by GQA [6].

The result is shown in Fig 2. As reported in [3], other methods that use spatial features such as BottomUp [1] and MAC [2] get an accuracy of 43%. In contrary, the attention map we use in LLaVA can achieve a higher accuracy, especially after layer 8. This indicates that using the attention map to select

[6] https://nlp.stanford.edu/data/gqa/eval.zip

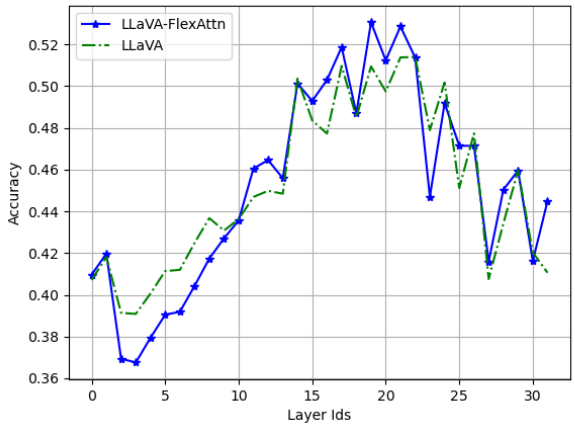

Fig. 2: Attention map accuracy analysis.

important regions is reasonable. Also, LLaVA equipped with FlexAttention has a slightly more accurate attention map than the vanilla model, meaning that the finetuning process can guide the model to pay more attention to the important regions.

A.4 Implmentation Details

High-resolution Image Encoder. We follow LLaVA-HD [7] to encode the high-resolution image by splitting the image into several subimages such that each subimage has the resolution that fits LLaVA's original vision encoder's resolution (*i.e.*, 336x336). We then encode each subimage independently using the original vision encoder and merge the feature patches of each subimage to form the patches for the high-resolution image.

Hyperparameters. We replace the original self-attention by our novel FlexAttention from layer 8 as we empirically find that the attention map before layer 8 does not highly correlated with the important regions. We hypothesize that it is because in lower layers the model is capturing some low level details such as edges and colors, and the useful semantic information only comes in later layers. We also set the attention map binarization threshold to be 0.1875 (*i.e.*, patches that has a normalized attention value higher than the threshold will be selected) to control the selected region ratio to be lower than 10%.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 **2**(4), 8 (2017)
2. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. arXiv preprint arXiv:1803.03067 (2018)
3. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
5. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. arXiv preprint arXiv:2311.15826 (2023)
6. Li, B., Zhang, P., Yang, J., Zhang, Y., Pu, F., Liu, Z.: Otterhd: A high-resolution multi-modality model. arXiv (2023)
7. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
8. Lobry, S., Marcos, D., Murray, J., Tuia, D.: Rsvqa: Visual question answering for remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **58**(12), 8555–8566 (2020)
9. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
10. Wu, P., Xie, S.: V*: Guided visual search as a core mechanism in multimodal llms. arXiv preprint arXiv:2312.14135 **17** (2023)