Identifying and Mitigating Position Bias of Multi-image Vision-Language Models

Supplementary Material

A. Experimental Details

Here, we provide a detailed description of the models, benchmarks, and evaluation details employed for this work.

Models. We provide an overview of each LVLM along with our specifications.

1) **OpenFlamingo** [2] represents a pioneering effort in the realm of multi-image LVLMs, serving as the open-source reproduction of Flamingo [1]. It maintains an almost identical architecture and training framework to the original. For our evaluation, we utilize the second version, OFv2, and conduct experiments on a model with 9 billion parameters.

2) **Idefics2** [8] is an 8-billion-parameter multi-image model pre-trained on newly collected interleaved web documents [7] and instruction datasets [8] from Huggingface. For our experiments, we adhere to its default settings, with the exception of disabling sub-image splitting for fairness.

3) **LLaVA-NeXT-Interleave** [9] is a recently developed model designed to enhance the multi-image, multi-frame, and multi-view reasoning capabilities of the LLaVA family. In this study, we utilize its 7-billion-parameter variant.

4) **InternVL2** [3] is a flexible multi-modal model that utilizes progressive alignment with large language models, ranging from 1 billion to 108 billion parameters. Here, we use its 8-billion-parameter variant, which employs dynamic resolution.

5) **VILA** [10] is a meticulously designed model that employs an empirically optimized pre-training strategy by reblending interleaved and image-text pair data. In this study, we utilize its 13-billion-parameter variant with 8-bit quantization.

6) **Mantis** [6] is an 8-billion-parameter state-of-the-art family enhanced from existing models through multi-image instruction tuning. Here, we select Mantis-Idefics2, which is directly fine-tuned from Idefics2 [8].

Benchmarks. We introduce the selected benchmarks here. In §3.1 and §5.1, to explore the impact of changing positions on predictions, we define position-agnostic tasks, which refer to tasks where changing the position does not alter the semantics of the question. Here, we list the position-agnostic tasks for each benchmark.

1) **BLINK** [5] is a novel benchmark comprising 14 tasks that can be effortlessly solved by humans yet present challenges for LVLMs.

2) **MuirBench** [12] is a comprehensive benchmark consisting of 12 tasks designed to evaluate the multi-image understanding capabilities of existing LVLMs.

3) **MIRB** [13] is a multi-image benchmark developed to assess the relational reasoning capabilities of LVLMs.

4) **Mantis-Eval** [6] is a compact test set designed for the Mantis family. Additionally, for comprehensiveness, we include some samples from Mantis-Instruct and ensure that they do not overlap with other benchmarks.

Evaluation. We uniformly employ LLM-as-a-judge for evaluation throughout the study. Specifically, we utilize GPT-4, providing it with the question, ground truth, and model prediction for each example, and instructing it to assess the correctness of the response. To maintain evaluation consistency, we use the default system prompt and set the temperature to 0, thereby employing greedy search decoding.

B. Limitations and Failure Cases

High sensitivety to hyperparameters. The proposed SoFA effectively mitigates position bias, yet its performance is highly sensitive to the choice of σ, which necessitates a small labeled validation set for calibration (32-shot in this study). Through our ablation study, we empirically demonstrate that no single global optimal σ is suitable across all tasks; instead, the optimal choice of σ is task-dependent and varies according to the specific characteristics of the task at hand. In future work, we could explore learning σ through gradient-based methods or directly fine tune the attention mask. Additionally, designing separate σ for each layer may be considered for better performance.

Limited to autoregressive models. SoFA is applicable to the vast majority of LVLMs that employ an autoregressive framework, wherein visual tokens are concatenated with text tokens and processed through the input layer. However, SoFA may not be suitable for certain specialized or custom architectures. For example, Flamingo [1] adopts a different approach by using image-text cross-attention, where text functions as the query and the image as both the key and value for interaction. Given the distinct nature of this architecture, we have not addressed it in this work and consider it as part of future research.

C. Societal Broader Impact

Our work has positive broader impacts. Intuitively, the position bias in LVLMs could be considered an imitation of human psychology, where people tend to selectively remember certain images while forgetting others while faced with multiple ones, a phenomenon called serial-position effect [4, 11]. We argue that this imbalanced reasoning capability across positions leads LVLMs to 1) potentially miss key information, impairing their performance and 2) produce unreliable responses influenced by image order. This is detrimental to various multi-image applications such as scene understanding and visual analogy. Our proposed SoFA, by slightly modifying attention, effectively mitigates this issue. Despite being an inference-only method, we hope our approach can also inspire the community to devise more remedies during the training stage.

Currently, we have not identified any significant negative impacts. However, this needs to be reassessed in the future due to objective factors such as the availability of datasets and models.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 1

[4] Hermann Ebbinghaus. A contribution to experimental psychology. *New York, NY: Teachers College, Columbia University*, 1913. 1

[5] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024. 1

[6] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 1

[7] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *NeurIPS*, 2024. 1

[8] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 1

[9] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 1

[10] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024. 1

[11] Bennet B Murdock Jr. The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482, 1962. 1

[12] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 1

[13] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024. 1