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

Xinyu Tian[1] Shu Zou[1] Zhaoyuan Yang[2] Jing Zhang[1]
[1]Australian National University [2]GE Research
[1]firstname.lastname@anu.edu.au, [2]firstname.lastname@ge.com

Abstract

The evolution of Large Vision-Language Models (LVLMs) has progressed from single to multi-image reasoning. Despite this advancement, our findings indicate that LVLMs struggle to robustly utilize information across multiple images, with predictions significantly affected by the alteration of image positions. To further explore this issue, we introduce Position-wise Question Answering (PQA), a meticulously designed task to quantify reasoning capabilities at each position. Our analysis reveals a pronounced position bias in LVLMs: open-source models excel in reasoning with images positioned later but underperform with those in the middle or at the beginning, while proprietary models show improved comprehension for images at the beginning and end but struggle with those in the middle. Motivated by this, we propose SoFt Attention (SoFA), a simple, training-free approach that mitigates this bias by employing linear interpolation between inter-image causal attention and bidirectional counterparts. Experimental results demonstrate that SoFA reduces position bias and enhances the reasoning performance of existing LVLMs. The code will be available https://github.com/xytian1008/sofa.

1. Introduction

The advent of Large Vision-Language Models (LVLMs) [1, 2, 5, 24, 34] has catalyzed the development of various text-grounded visual applications, including visual question answering [3, 16, 36], code generation [8, 37, 52], and story telling [13, 26]. Open-source LVLMs such as LLaVA [34], BLIP2 [28], and InternVL [11] have demonstrated competitive performance on single-image tasks like VQAv2 [16], OKVQA [38], and MMMU [64]. However, these LVLMs are limited in their capacity to analyze inter-image differences and relationships, as well as to process temporal information. Consequently, recent LVLMs, such as Flamingo [2, 4], Idefics [22–24], and Emu2 [51], have incorporated image-text interleaved data during training to enable ability to tackle multi-image tasks [15, 19, 50, 56].

Despite the increasing scale and performance of LVLMs, the interpretation of how they reason across multiple images remains a mystery. In fact, we might find some hints in the topic of natural language processing (NLP). Recent research in NLP has highlighted retrieval-augmented generation (RAG) [25, 44, 71], which aims to enable large language models (LLMs) to reason across multiple documents to answer questions. However, [35] find that LLMs tend to effectively comprehend documents positioned at the beginning and end of the input, often neglecting the middle sections. Similarly, in LLM-as-a-judge [48, 55], [68] find that LLMs used to evaluate two model responses tend to favor the first one, regardless of content. This phenomenon, where the model's reasoning capability is influenced by the segment position, is referred to as position bias [35, 42, 60].

Motivated by the research above, a natural question arises: does position bias exist in LVLMs? Specifically, we seek to determine whether LVLMs can robustly make use of information across multiple images. To investigate this, we begin by evaluating LVLMs on established multi-image benchmarks [15, 19, 50, 56]. We find that simply changing the order of presented images may significantly alter the predictions for most LVLMs, leading to considerable variation in benchmark accuracy. Fig. 1 provides examples of questions and model responses. For instance, in difference spotting [56], swapping two images causes OpenFlamingo [4] to miss distinctions, while in a simple counting task [19], reordering images leads Idefics2 [24] to incorrect reasoning. This positional sensitivity is apparent even in GPT-4o [1], a leading proprietary model.

The above motivational study reveals an unfortunate fact: building upon LLMs, LVLMs may inherit the notorious position bias from their language component. This bias manifests as the model's tendency to excel at reasoning visual information from certain positions, while exhibiting weaknesses at others, compromising robustness and reliability in predictions. To further investigate which positions display strong understanding and which reveal weaknesses, *i.e.*, identifying the pattern of such position bias, we introduce Position-wise Question Answering (PQA). Unlike most multi-image tasks that assess holistic comprehension, PQA requires LVLMs to reason consistently across each

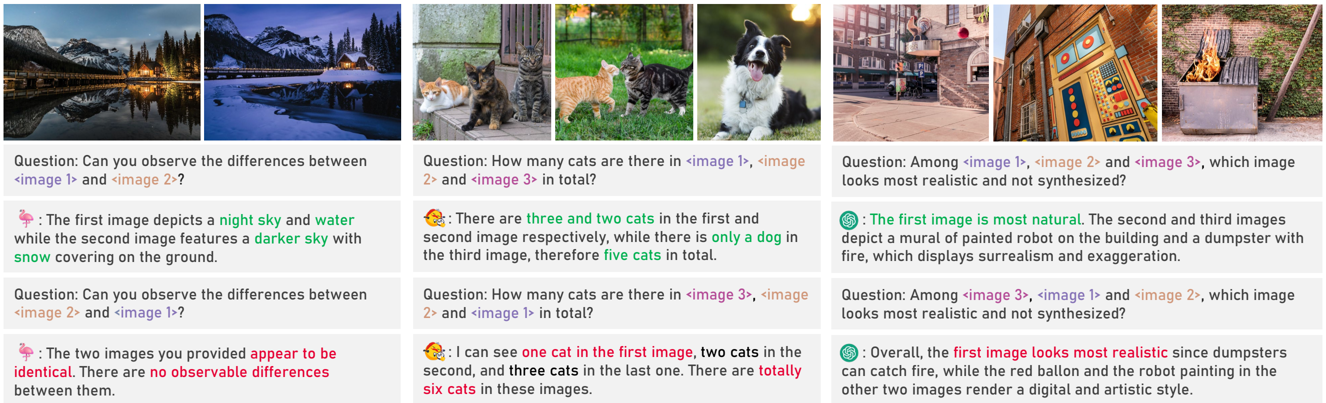


Figure 1. Examples where the predictions of LVLMs are influenced by changes in the image positions (`Date accessed: Oct 25, 2024`). The tasks, from left to right, are difference spotting [56], counting [19] and forensic detection [15], respectively. The correct answers are highlighted in green, while the incorrect ones are highlighted in red. The prompts and outputs are simplified here for clarity and aesthetics.

image and produce position-wise responses. This approach enables us to measure position-wise accuracy and analyze the model's reasoning capacity across different positions.

Our analysis reveals that most open-source models, *e.g.*, OpenFlamingo [4] and Idefics2 [24], exhibit a pronounced recency bias, demonstrating stronger reasoning capabilities for images presented later while underperforming on those in the middle or at the beginning. In contrast, proprietary models like GPT-4o [1] show enhanced understanding of images at the beginning and end, while neglecting those in the middle, a pattern consistent with previous findings in LLMs [35]. This observation is further corroborated by Fig. 1, where LVLMs tend to reason incorrectly about images positioned in poor-performing areas.

The existence of position bias significantly compromises model robustness and undermines accuracy across various multi-image tasks. Inspired by this, we aim to investigate the underlying mechanical causes and mitigate this issue. Specifically, we notice that in most LVLMs, the hidden states of image tokens are heavily influenced by their relative positions due to inter-image causal attention. In particular, image tokens are treated as a type of special text tokens where 1) images appearing later may interact with preceding ones; 2) images at the beginning are isolated, lacking access to global context. This autoregressive setup implicitly imposes a semantic order on the presented images, amplifying their positional information. To address this, we introduce SoFt Attention (SoFA), a training-free approach to mitigate position bias across various LVLMs. SoFA reduces the dependency of image hidden states on positional information by simply performing linear interpolation between inter-image causal attention and its bidirectional counterparts. Experimental results indicate that SoFA effectively equilibrates the reasoning capabilities of LVLMs across various positions, achieving average accuracy improvements of $2 \sim 3\%$ on existing multi-image benchmarks.

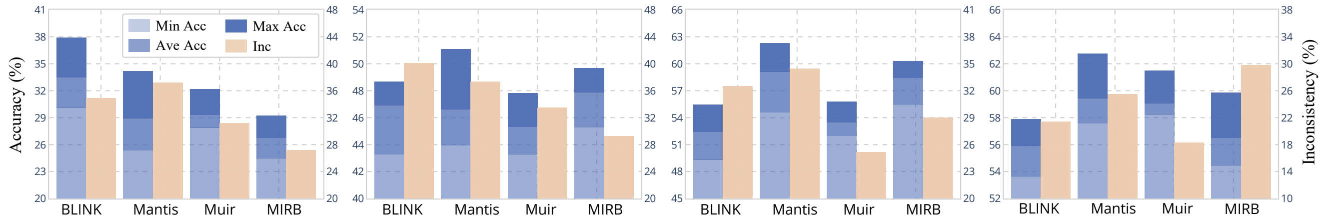
In summary, our contributions are as follows:

- Despite the significant potential of LVLMs in multi-image applications, we find that position bias, a well-explored phenomenon in NLP, has emerged as a potential bottleneck in the development of LVLMs, substantially impacting model robustness and reliability.
- We further identify such position bias and discover that most open-source models demonstrate a pronounced recency bias, whereas proprietary models tend to struggle in middle positions.
- We introduce SoFA, a straightforward training-free approach designed to mitigate the position bias inherited from pre-trained LVLMs, and substantiate its effectiveness through experimental validation.

2. Related Work

Large Vision-Language Models. Existing LVLMs are generally composed of a vision encoder [43, 65], an LLM [12, 53, 54], and a bridging projector [2, 7, 27] that facilitates interaction between the two. Conventional LVLMs, pre-trained on image-text pairs [46] and fine-tuned through preference [33, 40] or instruction tuning [27, 41], demonstrate substantial success in applications like dialogue [5, 11], question answering [58, 69], and domain-specific tasks [39, 47]. However, constrained by their training data, these models primarily support single-image reasoning during inference. Recently, the integration of image-text interleaved data [23, 70] has enabled multi-image reasoning capabilities, encompassing image comparison [21, 56], association [15, 19], temporal relationship analysis [29], and emerging properties of multi-modal in-context learning [30, 67]. Nonetheless, limited research has delved into how LVLMs perform reasoning across multiple images or the underlying mechanisms driving this process.

Position Bias in LLMs. Position bias, an inherent side effect of long-context input, has increasingly captured the at-



(A) OpenFlamingo (B) Idefics2 (C) LLaVA-NeXT (D) GPT-4o

Figure 2. The results of multiple evaluations while shuffling image positions. We record the minimum, maximum, and average accuracy (left vertical axis), along with the prediction inconsistency between the best and worst-performing evaluations (right vertical axis).

tention of NLP researchers. The pioneering work on this issue, notably [35], reveals that when LLMs process multiple documents, they often concentrate on those at the beginning and end while disregarding much of the information in between, *i.e.*, a phenomenon termed "lost in the middle". A related issue has also been identified in LLM-as-a-judge [48, 55], where [68] observe that LLMs tasked with evaluating multiple model responses tend to favor the first one, irrespective of its content. Recent studies have begun to explore methods to mitigate or circumvent position bias in LLMs [42, 45, 60, 63]. However, to the best of our knowledge, few studies have examined the effects of position bias within multi-modal applications, specifically regarding its implications for LVLMs.

Position Bias in LVLMs. While systematic research on position bias in LVLMs is still lacking, recent studies have indirectly underscored the influence of positions on LVLMs through various findings. For instance, in multi-modal in-context learning, studies [9] and [6] reveal that demonstration images placed at the beginning contribute minimally to LVLM performance. Additionally, in assessing long-context capabilities, [61] reports that LVLMs attain particularly high accuracy when key information appears at the end of the input sequence, with similar results observed in [59] and [57]. However, previous research has neither thoroughly analyzed this position bias nor proposed solutions to address it. In contrast, this study systematically investigates position bias in both open-source and proprietary models and introduces an effective mitigation strategy.

3. Understanding Position Bias in LVLMs

In this section, we investigate whether the position bias observed in LLMs extends to LVLMs. We start by empirically confirming the existence of position bias in LVLMs through a motivational study, revealing their poor robustness to image position changes (§3.1). Subsequently, we analyze the nature of this bias by evaluating and quantifying the reasoning capabilities of LVLMs regarding different positions (§3.2). These insights contribute to a deeper understanding and interpretation of LVLMs' reasoning behavior and common failure cases.

3.1. The Position Matters

To assess the existence of position bias, we conduct a simple motivational experiment, as shown in Fig. 2.

Experimental Setup. We select four representative multi-image LVLMs: OpenFlamingo [4], Idefics2 [24], LLaVA-NeXT-Interleave[1] [27], GPT-4o [1], and conduct evaluations across established multi-image benchmarks, including BLINK [15], Mantis-Eval [19], MuirBench [56], and MIRB [66]. To ensure clarity, we select partial tasks from these benchmarks that feature position-agnostic questions, wherein the positions of the images do not affect the answers to the questions. See §5.1 for the selected tasks and details. To explore the influence of image positions on the results, we perform multiple evaluations per benchmark, with the positions of the images shuffled in each iteration.

Metric. We record the minimum, maximum, and average accuracy across these evaluations, which directly reflects the influence of positions on the performance of LVLMs. Furthermore, we assess the prediction inconsistency between the best and worst-performing evaluations, quantifying the proportion of examples exhibiting conflicting predictions across the two evaluations by changing positions.

Results. As illustrated in Fig. 2, for most open-source LVLMs, the prediction inconsistency between the best and worst-performing evaluations approaches 30%, suggesting that merely altering the positions of the presented images may result in an amount of the model's predictions being changed. Such pronounced sensitivity to positional information results in highly unstable performance. For example, OpenFlamingo's accuracy on BLINK can vary significantly, reaching 38% at its peak and dropping to 30% at its lowest, indicating an accuracy span of up to 10% with position changes. GPT-4o, the state-of-the-art proprietary model, although showcasing superior performance, also displays a lack of robustness to positional variations, leading to a prediction inconsistency of around 25%.

To qualitatively explore this issue, we present examples in Fig. 1 that illustrate prediction conflicts arising from alterations in image positions. For instance, when prompted to identify the differences between two images, Open-

[1]For brevity, we refer to LLaVA-NeXT-Interleave as LLaVA-NeXT.

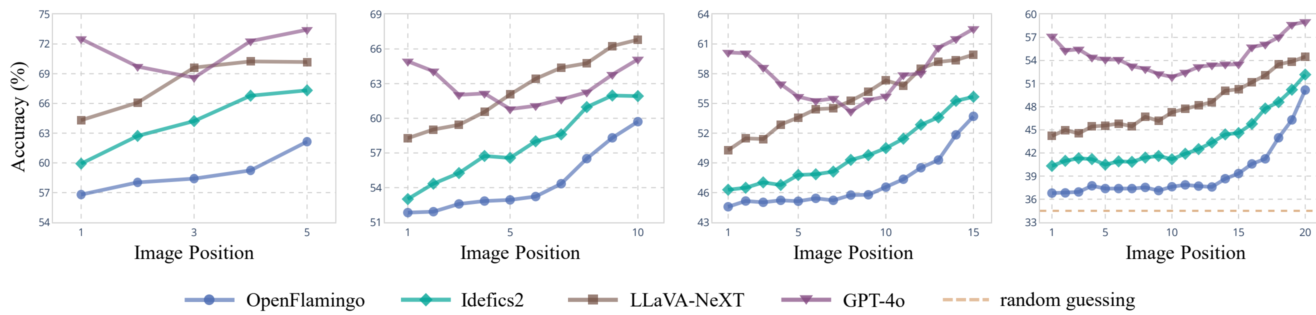


Figure 3. The results on the PQA task. We report position-wise accuracy on four scenarios where the number of images is 5, 10, 15 and 20, respectively. A higher accuracy signifies strong reasoning at the specified position, whereas lower reflects poor-performing areas.

Flamingo fails to extract fine-grained distinctions upon swapping their positions. Similarly, in the task of object counting, Idefics2 mistakenly identifies a dog as a cat due to the repositioning of images. GPT-4o demonstrates a comparable issue, where rearranging the image options alters both its answer and the accompanying rationale.

3.2. Identifying Position Bias

The quantitative results and qualitative examples above imply biased reasoning across positions from LVLMs. To have a closer look at such position bias, we aim to identify models' reasoning capabilities at different image positions.

However, existing tasks have been restricted to evaluating models' holistic comprehension of multiple images, lacking the means to systematically analyze their fine-grained understanding of each individual position. Specifically, these tasks provide coarse ground truth answers, such as option symbols or numeric values, but do not offer individual rationales for each image, making it challenging to quantify the model's reasoning across different positions. To address this gap, we introduce Position-wise Question Answering (PQA), a carefully designed task that assesses the position-wise reasoning capabilities of LVLMs. In PQA, each question prompts LVLMs to reason independently across all images, generating a distinct response for each. This approach 1) ensures task-level neutrality with respect to any particular position and 2) enables quantification of LVLMs' reasoning performance for each position.

Specifically, taking object counting as an example, we employ the following query template:

```
I: Please provide one answer for each
    image in the form like [x, x, x, ...].
Q: Among <image 1>, ..., <image N>, how
    many cats can you find in these images?
A: [3, 2, 0, ...].
```

Each query consists of an instruction (`I`) and a question (`Q`), where the former prompts the model to generate position-wise responses (`A`) following a specified output format. We gather responses from LVLMs as a structured list, with each element representing a particular position. This enables us to track position-wise accuracy, where a higher accuracy reflects stronger model capability at a given position, while lower indicates areas of relative weakness.

Construction of position-wise examples. Since in PQA, each image is semantically independent of others, a PQA example essentially consists of several single-image instances that share the same question. This allows us to easily build position-wise examples from single-image benchmarks. Specifically, we select VQAv2 [16], a substantial single-image dataset that encompasses a diverse array of everyday topics and features both multiple-choice and open-ended questions. Initially, we identify and consolidate single-image examples with identical questions to create a question set, wherein each question corresponds to multiple images. Subsequently, we reformulate these questions into PQA format according to the above-mentioned template.

Experimental Setup. We utilize the same LVLMs as previously mentioned and conduct evaluations on the proposed PQA task. Additionally, we also consider random guessing, where binary choice and numeric questions are answered randomly with yes or no, or with an integer value. We examine four scenarios with varying numbers of images: 5, 10, 15, and 20, collecting 1,000 examples for each scenario. Furthermore, to maintain fairness, we augment each example by shuffling the image positions four times, resulting in a total of 5,000 examples. In contrast to prior settings that evaluate holistic performance, we provide separate accuracy metrics for each position in this analysis.

Results. Fig. 3 displays the results of LVLMs on the PQA task, from which we may derive the following insights:

1) **Open-source LVLMs exhibit recency bias, whereas proprietary models struggle in middle positions**. For open-source models, a stronger reasoning capability is observed for images positioned at the end, while significant underperformance is noted in the middle and beginning. For example, with 10 images, the accuracy of OpenFlamingo varies by nearly 9% between the far left and far right positions (51.84% $\rightarrow$ 59.71%). In contrast, proprietary models, *i.e.*, GPT-4o, reveal a typical U-shaped curve, highlighting pronounced weaknesses in the middle positions.

2) **Position bias becomes more pronounced as the number of images increases**. For instance, with 20 images, the accuracy gap for OpenFlamingo between the two ends reaches 14% (36.79% $\rightarrow$ 50.17%). Notably, at this stage, OpenFlamingo seems to nearly lost its reasoning capability for most of the beginning images, achieving results only marginally better than random guessing (34.49%).

3) **Relationship to position bias in LLMs**. Previous research [35] shows that LLMs often overlook information in middle positions when processing multiple documents, a phenomenon termed "lost in the middle". Our findings suggest this effect persists for GPT in multi-image reasoning. However, for open-source models, the primacy position advantage appears to diminish, showing a marked decaying trend from the end to beginning. Whether the position bias in LVLMs is correlated with that of their LLM backbones remains an interesting question for further exploration.

4. The Proposed Method

In this section, we first analyze the underlying causes of position bias (§4.1). Then we introduce SoFA, a straightforward, training-free approach to mitigate this issue (§4.2).

4.1. Position Mechanism in LVLMs

To investigate the origins of position bias, we need to first determine what contributes to the positional information in LVLMs. In NLP, it is generally believed that position embeddings [45, 49] and causal attention [42, 60] are the primary mechanisms enabling the sequential structure of input. Formally,

$$\begin{aligned} \mathbf{Q} &= \mathcal{P}(\mathbf{X}\mathbf{W_q}, \mathbf{pos}), \quad \mathbf{K} = \mathcal{P}(\mathbf{X}\mathbf{W_k}, \mathbf{pos}) \\ \mathbf{H} &= \mathrm{Softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}}/\sqrt{d}) \odot \mathbb{1}_{\text{causal}}\mathbf{V} \end{aligned} \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the queries, keys, and values, respectively. $\mathbf{X}$ refers to the input tokens, encompassing both image and text types, $\mathbf{pos}$ denotes the position embedding of $\mathbf{X}$, which is integrated into the attention via the position encoding function $\mathcal{P}$, $\mathbb{1}_{\text{causal}}$ denotes the causal mask, which ensures autoregressive generation.

This reveals two key factors influencing LVLMs' positional awareness: 1) position embedding explicitly initializes a sequential value to each token, and 2) causal attention implicitly enhances this positional characteristic through unidirectional information access. This presents us with two distinct solutions. One approach is to directly reassign position embeddings, a widely used strategy in NLP [10, 18, 45], as it fundamentally eliminates the influence of position on documents. However, this aggressive method is not suitable for LVLMs, since in many scenarios, such as video understanding and action recognition, we still expect images to retain positional information for temporal analysis. To clarify, our aim is to mitigate the imbalanced

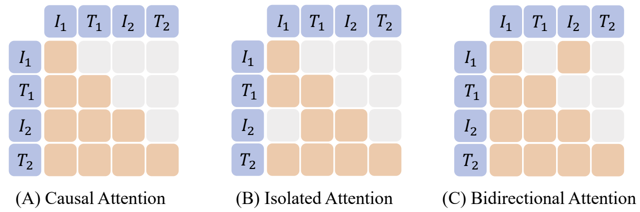


Figure 4. The three inter-image attention mechanisms, where I_1 and I_2 represent the tokens of two images, T_1 and T_2 represent the tokens of two text segments. Note that typically each image involves numerous tokens, *e.g.*, 576 for LLaVA. Here for clarity we simplify them to a single token. In (A), images interact in a unidirectional manner, allowing I_2 to attend to I_1, while I_1 remains isolated. In (B), each image is isolated, indicating they can only attend to themselves. In (C), bidirectional interaction is enabled so that each image can attend to any other images. It is worth mentioning that we only alter the inter-image attention while preserving causal attention between the text segments.

reasoning capabilities of LVLMs across different positions, rather than to eliminate their positional awareness.

Therefore, in this paper, we primarily investigate the relationship between causal attention and position bias. We notice that, while images do not exhibit a distinct semantic order as text does, most LVLMs, *e.g.*, LLaVA series, treat image and text tokens in a uniform autoregressive manner. As shown in Fig. 4 (A), analogous to texts, there exists inter-image causal attention which allows 1) later-positioned images to interact with those preceding them, while 2) earlier images remain isolated, lacking access to global contexts. From a mechanical perspective, this framework renders image hidden states highly dependent on their relative positions, thus implicitly amplifying their positional cues.

To verify our hypothesis, we compare the inter-image causal attention against following two alternative variants:

1) **Isolated attention**. As illustrated in Fig. 4 (B), we disable interactions between images, enforcing each image to attend solely to itself.

2) **Bidirectional attention**. Rather than restricting to unidirectional information flow, we permit each image to attend to any other, as shown in Fig. 4 (C).

We modify the default causal mask in the language decoder to its isolated and bidirectional counterparts and assess the impact of them on position bias in PQA, as shown in Fig. 5. It is evident that altering causal attention to either isolated or bidirectional forms significantly alleviates position bias, resulting in more balanced accuracy across different positions. This empirically confirms that inter-image causal attention is the primary source of position bias. However, these modifications lead to varying degrees of performance degradation, with isolated attention causing the most substantial decline. This outcome is reasonable, as such modifications may substantially deviate from the training framework, leading to out-of-distribution hidden states.

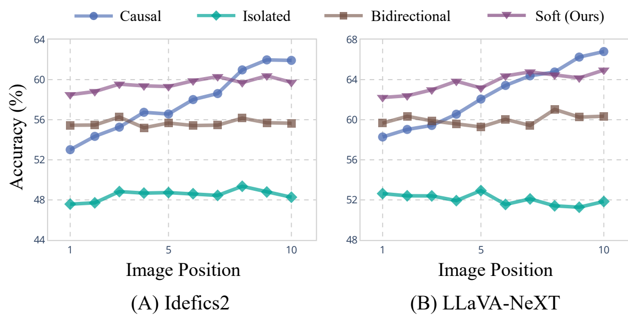


Figure 5. The position-wise accuracy of the three attention mechanisms on the PQA task, alongside our proposed SoFA method.

4.2. Mitigating Position Bias

The analysis above shows that inter-image causal attention is the main cause of position bias. Furthermore, it also reveals a trade-off between accuracy and robustness, a typical dilemma in machine learning. Building on this insight, we propose SoFt Attention (SoFA), a straightforward, training-free approach to mitigate position bias. SoFA employs linear interpolation between inter-image causal attention and its bidirectional counterpart. Formally,

$$\begin{aligned} \mathbf{H}_{\text{soft}} &= \text{Softmax}(\mathbf{Q}\mathbf{K}^{\text{T}}/\sqrt{d}) \odot \mathbf{M}_{\text{soft}}\mathbf{V} \\ \mathbf{M}_{\text{soft}} &= (1-\sigma)\mathbb{1}_{\text{causal}} + \sigma\mathbb{1}_{\text{bidirectional}} \end{aligned} \tag{2}$$

where $\mathbb{1}_{\text{causal}}$ and $\mathbb{1}_{\text{bidirectional}}$ denote the two attention masks, with σ controlling the extent of bidirectional part. Note that since we are only interested in the visual positional relation, we only modify the inter-image attention, while preserving causal attention between text segments.

SoFA as a position smoother. Intuitively, given that inter-image causal attention excessively emphasizes positional information, SoFA integrates partial bidirectional attention to smooth the influence of position on the models at minimal cost, achieving a balanced trade-off between reasoning performance and position robustness.

SoFA as an interaction enhancer. Practically, SoFA enhances the interaction between images, enriching the semantic information of earlier image features to enable more effective utilization by LVLMs. This aligns with the most cases where key information is typically distributed evenly across various input positions.

For each task, we utilize a small validation set to determine the optimal values for σ. Besides, to better align with the training framework, we deploy SoFA every two layers rather than at every layer, while preserving causal attention in the remaining layers. We present SoFA's results in the PQA task as shown in Fig. 5. It is evident that SoFA significantly outperforms both isolated and bidirectional attention, while unlike causal attention in the default setting, SoFA demonstrates balanced performance and exhibits enhanced reasoning capability in most positions.

5. Experiment and Results

In this section, we empirically demonstrate that SoFA 1) effectively mitigates the position bias inherent in LVLMs and 2) significantly improves their overall reasoning capability across various multi-image applications.

Models. We select a diverse set of open-source models, including Idefics2-8B [24], InternVL2-8B [11], VILA-13B [31], Mantis-8B [19], and LLaVA-NeXT-Interleave-7B [27].

Implementation details. For the input, we provide instructions that explicitly indicate the task type and output formats. For instance, for multiple-choice questions, we ask LVLMs to provide only the option symbol; for numeric questions, only the number; and for open-ended questions, the response is limited to short answers. For evaluation, to ensure fairness, we uniformly adopt GPT-4 [1], providing with the question, ground truth answer, candidate prediction, and it determines whether the prediction is correct or not. For SoFA, we determine σ using a 32-shot validation set per task and perform interpolation every two layers. Unless otherwise specified, all models are set to FP16 precision and utilize Flash Attention [14], with sub-image splitting [32] disabled for fairness.

5.1. Position Robustness

Benchmark. We first evaluate whether SoFA may effectively mitigate position bias. To this end, similar to §3.1, we select position-agnostic tasks from existing multi-image benchmarks. Specifically, we exclude tasks where changing the order of images may alter the semantics of the question, rendering it unanswerable, *e.g.*, video understanding. Here, we list some of the selected tasks: forensic detection and image jigsaw from BLINK [15], difference spotting and diagram understanding from MuirBench [56], visual analogy and attribute matching from MIRB [66]. We include all examples from Mantis-Eval [19] since the majority are position-agnostic. See the supplementary material for the details of more evaluated tasks. Our experimental setup and metrics are consistent with those described in §3.1.

SoFA reduces the influence of positions on LVLMs. As demonstrated in Table 1, the integration of SoFA within the models markedly reduces prediction inconsistency, leading to enhanced stability in performance. For instance, Idefics2's inconsistency on BLINK declines from 41.55% to 12.36%, signifying that the responses for the majority of examples remain consistent despite variations in image positioning. This observation indicates that LVLMs exhibit more balanced reasoning capabilities across different positions, effectively mitigating the effects of position bias.

SoFA eliminates the blind spots of LVLMs. In addition to the enhanced robustness, we observe a marked improvement in the performance of LVLMs in worst-case scenarios, particularly regarding minimum and average accuracy,

Method	BLINK [15]				Mantis-Eval [19]				MuirBench [56]				MIRB [66]			
	Min	Ave	Max	Inc	Min	Ave	Max	Inc	Min	Ave	Max	Inc	Min	Ave	Max	Inc
Idefics2 [24]	43.26	46.93	48.68	41.55	43.93	46.62	51.10	38.57	43.28	45.30	47.81	34.46	45.26	47.88	49.68	29.94
Idefics2 + SoFA	47.18	48.63	49.12	12.36	48.71	49.18	50.83	**8.61**	46.12	47.51	48.45	13.27	47.39	48.36	49.13	6.99
InternVL2 [11]	38.81	40.45	43.74	30.18	45.36	46.92	49.51	25.24	49.28	52.33	56.10	38.65	41.84	44.38	46.36	29.60
InternVL2 + SoFA	41.64	42.32	43.51	**7.16**	48.13	49.25	50.11	12.00	**54.69**	**55.92**	**56.78**	**5.16**	44.20	45.35	45.87	**6.64**
VILA [31]	45.93	48.59	51.42	25.33	48.44	49.20	51.85	21.29	41.66	43.12	48.27	37.26	47.17	49.59	52.34	31.77
VILA + SoFA	48.27	50.80	51.17	10.68	49.29	51.60	52.68	12.16	45.76	46.52	47.13	7.92	48.22	51.43	51.95	17.46
Mantis [19]	48.34	49.24	51.52	28.35	56.40	58.38	63.42	32.12	47.67	48.94	51.15	27.45	53.11	55.71	57.42	25.79
Mantis + SoFA	49.22	50.87	52.34	16.23	**60.48**	**62.21**	**64.68**	14.30	49.88	50.26	50.79	5.61	54.39	56.34	56.93	8.35
LLaVA-NeXT [27]	49.34	52.40	55.43	31.89	54.57	59.10	62.29	33.75	51.96	53.45	55.78	24.92	55.49	58.40	60.28	28.56
LLaVA-NeXT + SoFA	**53.20**	**54.91**	**56.08**	11.20	58.87	61.23	61.98	14.20	53.61	55.37	56.49	9.15	**58.13**	**59.81**	**61.16**	6.96

Table 1. The evaluation on position-agnostic tasks with and without SoFA. Similar to §3.1, we perform multiple evaluations and record minimum, maximum and average accuracy. We also report prediction inconsistency between best and worst-performing evaluations.

Method	BLINK	M-Eval	Muir	MIRB	NLVR2	MVBench
Random Guessing	39.27	25.31	21.49	19.13	50.75	25.91
Idefics2 [24]	45.78	48.41	42.84	46.58	81.94	23.37
Idefics2 + SoFA	48.31	50.77	47.36	48.83	85.15	25.65
InternVL2 [11]	38.95	50.30	54.53	42.66	85.56	29.31
InternVL2 + SoFA	43.26	51.11	**57.14**	46.19	88.19	32.77
VILA [31]	49.79	52.64	48.73	50.36	78.82	25.81
VILA + SoFA	50.29	52.13	52.80	51.15	80.34	26.40
Mantis [19]	51.41	56.85	44.28	54.59	**90.26**	41.59
Mantis + SoFA	53.19	**59.23**	47.64	55.68	89.98	44.40
LLaVA-NeXT [27]	53.34	50.83	48.22	57.15	87.28	54.26
LLaVA-NeXT + SoFA	**55.92**	54.51	50.43	**60.67**	89.45	**57.71**

Table 2. The overall evaluation accuracy of LVLMs on various multi-image benchmarks with and without SoFA.

which both exhibit an increase of approximately 3%. This suggests that the model retains strong performance despite variations in the distribution of image positions, indicating that SoFA effectively addresses the deficiencies of LVLMs in positions where performance is typically suboptimal.

5.2. Overall Results

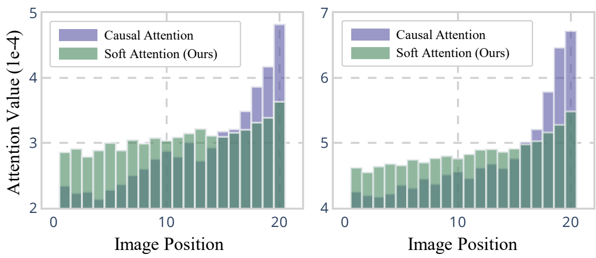
Benchmark. Given the observation that SoFA effectively mitigates position bias, we proceed with a comprehensive evaluation to determine whether SoFA can enhance the overall reasoning capabilities of LVLMs. Specifically, we include all tasks from the selected benchmarks [15, 19, 56, 66]. In addition, we incorporate NLVR2 [50] and MVBench [29], which exclusively assess the models' capability in image-pair comparison and video understanding, respectively. We directly utilize the default image positions from the benchmarks and record their accuracy.

SoFA enhances reasoning capability of LVLMs. In Table 2, we observe that integrating SoFA leads to an average improvement of $2 \sim 3\%$ in the performance of LVLMs across various benchmarks. For instance, Idefics2 shows a 2.25% improvement on MIRB ($46.58\% \rightarrow 48.83\%$), while InternVL2 exhibits a 4.31% increase on BLINK ($38.95\% \rightarrow 43.26\%$). This demonstrates that, despite the advancements in multi-image LVLMs, position bias poses a significant obstacle to their reasoning capability, whereas SoFA successfully addresses this limitation.

5.3. Discussion

Attention distribution visualization. To gain a deeper understanding of SoFA's working mechanism, we visualize the attention distribution within LVLMs across different positions. Using PQA as an example, we derive the average attention value of the final layer's query token (whose hidden state is directly used for generation) for each image. As shown in Fig. 6, under the default setting, the attention distribution is dominated by the end positions, indicating that the model overly focuses on a few images while neglecting most others. This observation aligns with our findings in §3.2. The application of SoFA alters this trend by significantly smoothing the attention distribution, enabling the model to attend to images at all positions more evenly. This well explains how SoFA mitigates position bias.

SoFA improves in-context capability of LVLMs. Multimodal in-context learning is an emerging property recently acquired by LVLMs, wherein a few demonstrations allow them to recognize task formats, thereby significantly improving accuracy. However, recent studies [6, 9, 30] find that LVLMs tend to overlook demonstration images that appear earlier, leading to suboptimal performance. Here, we demonstrate that SoFA can improve the in-context learning capability of LVLMs by eliminating such blind spots. Following [30], We select three typical visual question answering benchmarks: VQAv2 [16], VizWiz [17], and OK-VQA [38], and report the average accuracy of LVLMs across varying shot counts. As illustrated in Table 3, the application of SoFA leads to a further enhancement in LVLM performance, with this effect becoming increasingly pronounced as the number of shots rises. For instance, in the



(A) Idefics2 (B) LLaVA-NeXT

Figure 6. The attention distribution across positions on PQA.

Method	VQAv2 [16]			VizWiz [17]			OK-VQA [38]		
	4-shot	8-shot	16-shot	4-shot	8-shot	16-shot	4-shot	8-shot	16-shot
w/o SoFA	55.34	59.32	61.37	41.56	44.65	45.35	45.19	48.30	49.02
w/ SoFA	56.13	61.30	64.85	42.30	46.98	49.17	45.93	49.87	52.36

Table 3. The average accuracy of LVLMs on visual question answering using in-context learning with and without SoFA.

# Images	3	5	10	20	30	50	75	100
w/o SoFA	82.16	81.27	77.18	72.36	68.50	60.80	55.13	49.19
w/ SoFA	83.90	82.33	78.92	74.30	70.67	63.27	59.38	55.11

Table 4. The average accuracy of LVLMs on Multi-modal Haystack [59], where images range from a few to a hundred.

case of VizWiz, accuracy improves from 41.56% to 42.30% with 4 shots, and from 45.35% to 49.17% with 16 shots. This indicates that position bias undermines the effectiveness of demonstrations, while SoFA mitigates this issue.

SoFA exhibits advantages in long-context applications. Long-context scenarios have recently become a focus, where a single inference requires processing a substantial number of images. This is particularly beneficial for applications such as many-shot in-context learning [20] and multi-modal RAG [62]. Here, we show that SoFA can excel in these challenging scenarios. We select Visual Haystack [61], a benchmark designed to assess the long-context capabilities of LVLMs, and increase the input scale to encompass anywhere from a few images to a hundred. As shown in Table 4, the benefits of SoFA become increasingly evident as the number of images rises, achieving peak performance gains at 100 images ($49.19\% \rightarrow 55.11\%$). This indicates that an increase in context length exacerbates the model's position bias, while SoFA compensates this loss.

Contribution of SoFA to different tasks. Extending beyond the holistic evaluation, we categorize the benchmarks into various tasks to identify which ones benefit the most from SoFA. As depicted in Fig. 7, SoFA demonstrates the highest improvements in visual retrieval and analogy, with gains of 6.84% and 5.53%, respectively. In these tasks, the first image acts as a reference, and the models need to select a matching scene or similar style from the remaining images. This necessitates that the model pays extra attention

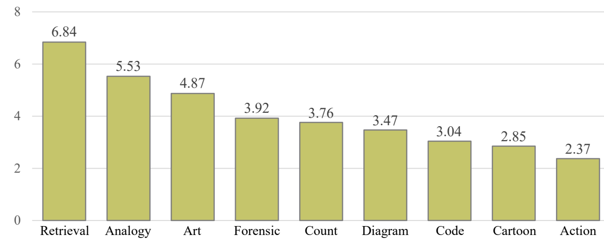


Figure 7. The performance gains of SoFA on different types of tasks. The results are averaged over selected models.

σ	BLINK	Mantis-Eval	MuirBench	MIRB	NLVR2	MVBench
0.00	47.85	51.81	47.72	50.27	84.77	34.87
0.25	49.22	52.70	49.35	**52.34**	85.41	**37.28**
0.50	**50.04**	53.33	**50.96**	51.86	86.30	36.56
0.75	49.62	**53.74**	48.24	51.14	**86.83**	35.07
1.00	48.21	51.58	47.37	50.53	85.37	34.20

Table 5. The results on six benchmarks varying the value of σ.

to and adequately understands the first image. Intuitively, due to the recency bias inherent in open-source models, they often overlook earlier information, resulting in suboptimal performance. Thus, the significant gains brought by SoFA are understandable.

The interpretation of σ. In Table 5, we conduct an ablation study on the effect of σ. It can be observed that the optimal σ varies significantly across benchmarks, for instance, with a value of 0.75 on NLVR2 and only 0.25 on MVBench. Intuitively, MVBench inputs involve video frames, which have a strong dependence on image positions, and an excessively high σ may compromise such positional information. In contrast, the examples in NLVR2 are image pairs, most of which are position-agnostic. This difference in task nature accounts for the variation in σ.

6. Conclusion

In this study, we introduce the concept of position bias, a well-known issue in NLP, into the realm of multi-modal applications. First, we reveal that LVLMs cannot robustly make use of information across multiple images, where merely altering the position of images may significantly change their predictions. Then we further identify this position bias, concluding that open-source models exhibit a strong recency bias, while proprietary models struggle with positions in the middle. Motivated by these findings, we propose SoFA, a simple, training-free approach that reduces the model's dependence on positional information through linear interpolation between inter-image causal attention and bidirectional counterparts. Experiments demonstrate that SoFA effectively mitigates position bias and significantly enhances the overall reasoning performance of existing models.

Acknowledgement

This research was, in part, funded by the U.S. Government – DARPA TIAMAT HR00112490421. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 3, 6

[2] 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

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *CVPR*, pages 2425–2433, 2015. 1

[4] 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, 2, 3

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1, 2

[6] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *CVPR*, pages 1539–1550, 2024. 3, 7

[7] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 2

[8] Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large language models are visual reasoning coordinators. In *NeurIPS*, 2024. 1

[9] Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models. In *ICLR*, 2023. 3, 7

[10] Yuhan Chen, Ang Lv, Ting-En Lin, Changyu Chen, Yuchuan Wu, Fei Huang, Yongbin Li, and Rui Yan. Fortify the shortest stave in attention: Enhancing context awareness of large language models for effective tool use. In *ACL*, 2023. 5

[11] 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, 2, 6, 7

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 2

[13] Neil Cohn, Ryan Taylor, and Kaitlin Pederson. A picture is worth more words over time: Multimodality and narrative structure across eight decades of american superhero comics. *Multimodal Communication*, 6(1):19–37, 2017. 1

[14] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*, pages 16344–16359, 2022. 6

[15] 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, 2, 3, 6, 7

[16] 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 *CVPR*, pages 6904–6913, 2017. 1, 4, 7, 8

[17] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 7, 8

[18] Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *ACL*, pages 13628–13642, 2024. 5

[19] 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, 2, 3, 6, 7

[20] Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y Ng. Many-shot in-context learning in multimodal foundation models. In *ICML*, 2024. 8

[21] Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Dee Guo, Sreenivas Gollapudi, et al. Remi: A dataset for reasoning with multiple images. *arXiv preprint arXiv:2406.09175*, 2024. 2

[22] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 1

[23] 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. 2

[24] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 1, 2, 3, 6, 7

[25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *CVPR*, pages 9459–9474, 2020. 1

[26] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, pages 13299–13308, 2024. 1

[27] 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. 2, 3, 6, 7

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 1

[29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024. 2, 7

[30] Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *CVPR*, pages 26710–26720, 2024. 2, 7

[31] 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. 6, 7

[32] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 6

[33] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 1

[35] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. In *ACL*, pages 157–173. MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA ..., 2024. 1, 2, 3, 5

[36] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pages 2507–2521, 2022. 1

[37] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. In *NeurIPS*, 2024. 1

[38] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204, 2019. 1, 7, 8

[39] Duy MH Nguyen, Hoang Nguyen, Nghiem Diep, Tan Ngoc Pham, Tri Cao, Binh Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. In *NeurIPS*, 2024. 2

[40] Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *ECCV*, 2024. 2

[41] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *ICLR*, 2023. 2

[42] Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023. 1, 3, 5

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2

[44] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. In *ACL*, pages 1316–1331. MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA ..., 2023. 1

[45] Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. Parallel context windows for large language models. In *ACL*, 2022. 3, 5

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2

[47] Junjie Shentu and Noura Al Moubayed. Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs. In *WACV*, pages 5212–5221, 2024. 2

[48] Lin Shi, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*, 2024. 1, 3

[49] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[50] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2018. 1, 7

[51] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, pages 14398–14409, 2024. 1

[52] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *CVPR*, pages 11888–11898, 2023. 1

[53] MosaicML NLP Team et al. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. 2

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[55] Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024. 1, 3

[56] 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, 2, 3, 6, 7

[57] Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.11230*, 2024. 3

[58] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2

[59] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. In *NeurIPS*, 2024. 3, 8

[60] Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. *arXiv preprint arXiv:2407.01100*, 2024. 1, 3, 5

[61] Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Visual haystacks: Answering harder questions about sets of images. *arXiv preprint arXiv:2407.13766*, 2024. 3, 8

[62] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *ICML*, 2022. 8

[63] Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. Mitigate position bias in large language models via scaling a single dimension. *arXiv preprint arXiv:2406.02536*, 2024. 3

[64] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024. 1

[65] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *CVPR*, pages 11975–11986, 2023. 2

[66] 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. 3, 6, 7

[67] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. In *ICLR*, 2023. 2

[68] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, pages 46595–46623, 2023. 1, 3

[69] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2023. 2

[70] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. In *NeurIPS*, 2024. 2

[71] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. In *NeurIPS*, pages 50117–50143, 2023. 1