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

Background

VLMs are evolving from single-image to multi-image reasoning

- LLaVA-1.5 → LLaVA-NeXT-Interleave, LLaVA-OneVision
- BLIP-2, InstructBLIP → X-InstructBLIP

LAION-5B

C: Green Apple Chair

C: sun snow dog

C: Color Palettes

C: pink, japan, aesthetic image

MMC4

[..., "Check out Shane Driscoll's take on sustainable communities and how his photograph fits this year's Green Cities theme.", ..., ,"Man-made platforms like the one pictured here allow these fish-eating birds of prey to thrive in developed coastal areas.", "A city surrounded by mountains.", "I took this photo in October on a hike in New Hampshire.", , "It is looking at Mt. Chicora from the middle sister mountain.", "Getting people out into beautiful places like this is becoming more and more popular, and each time we bring a little piece of nature back with us that inspires us to make our cities better.", ...]

[1] Schuhmann, Christoph, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models." *Advances in Neural Information Processing Systems* 35 (2022): 25278-25294.
[2] Zhu, Wanrong, et al. "Multimodal c4: An open, billion-scale corpus of images interleaved with text." *Advances in Neural Information Processing Systems* 36 (2024).

Background

The capability gap from single-image to multi-image reasoning

1. Cross-reasoning 2. Reference 3. Comparison 4. Temporal understanding

OK-VQA

Vehicles and Transportation

Q: What sort of vehicle uses this item?
A: firetruck

Brands, Companies and Products

Q: When was the soft drink company shown first created?
A: 1898

Objects, Material and Clothing

Q: What is the material used to make the vessels in this picture?

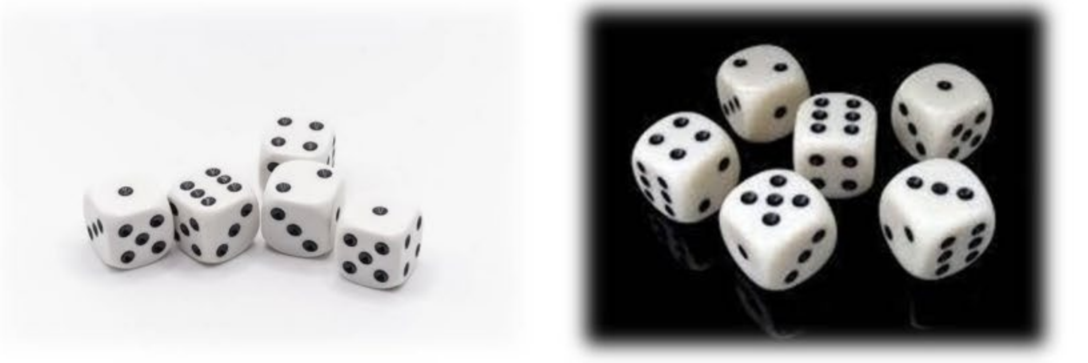
Sports and Recreation

Q: What is the sports position of the man in the orange shirt?
A: goalie

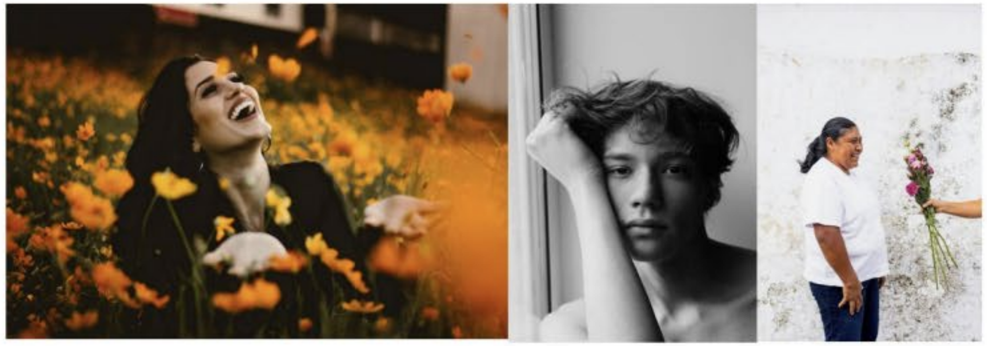
Cooking and Food

Q: What is the name of the object used to eat this food?
A: chopsticks

Mantis-Eval



How many dices are there in image 1 and image 2 respectively?



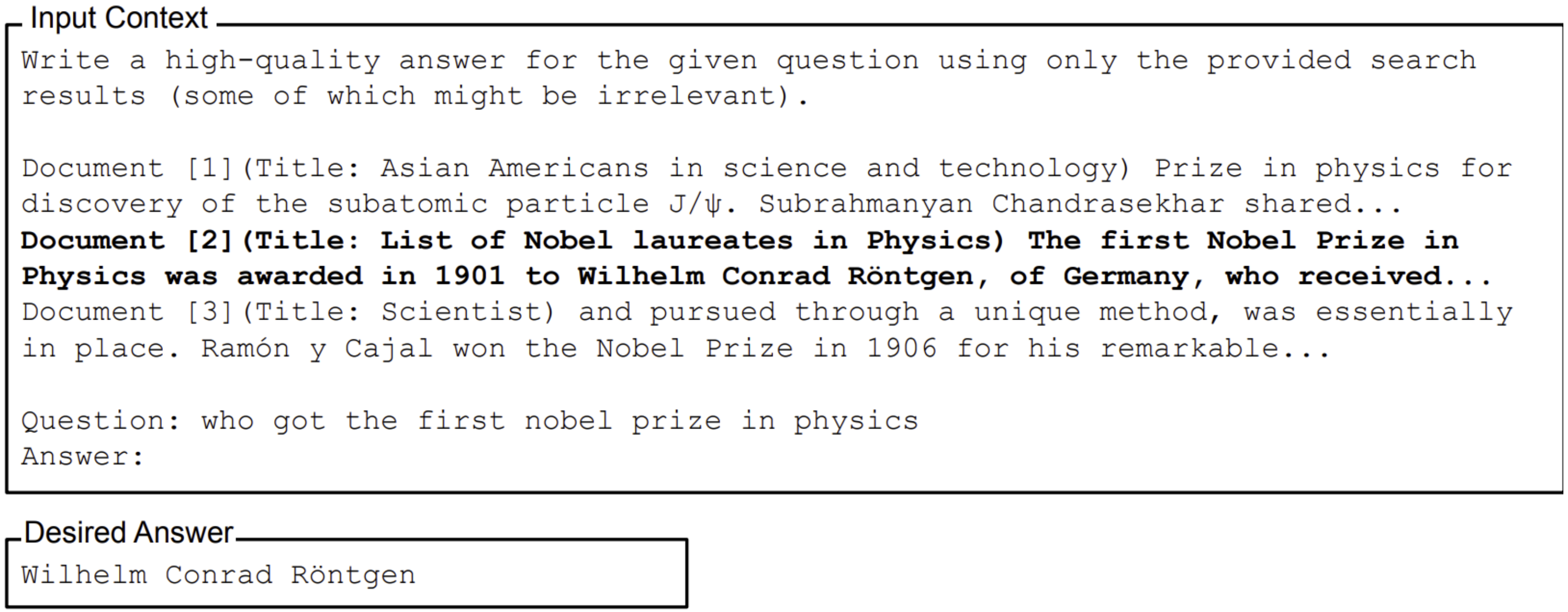
Which image shows a different mood of character from the others?

[3] Marino, Kenneth, et al. "Ok-vqa: A visual question answering benchmark requiring external knowledge." *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 2019.
[4] Jiang, Dongfu, et al. "Mantis: Interleaved multi-image instruction tuning." *arXiv preprint arXiv:2405.01483* (2024).

Motivation

Question: Does VLMs treat every image equally?

A similar dilemma in NLP where multiple documents are fed into LLMs



Input Context

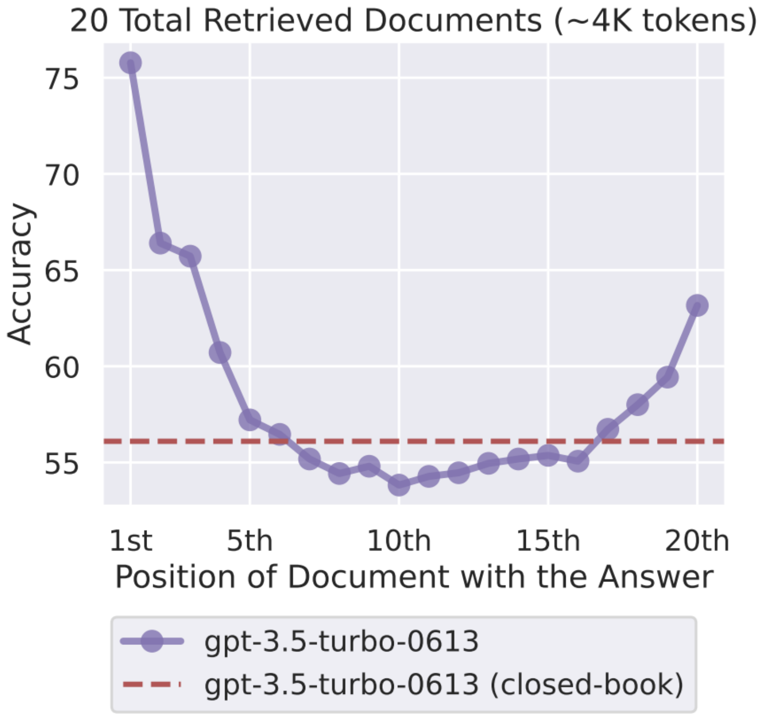
Write a high-quality answer for the given question using only the provided search results (some of which might be irrelevant).

Document [1](Title: Asian Americans in science and technology) Prize in physics for discovery of the subatomic particle J/ψ. Subrahmanyan Chandrasekhar shared...
Document [2](Title: List of Nobel laureates in Physics) The first Nobel Prize in Physics was awarded in 1901 to Wilhelm Conrad Röntgen, of Germany, who received...
Document [3](Title: Scientist) and pursued through a unique method, was essentially in place. Ramón y Cajal won the Nobel Prize in 1906 for his remarkable...

Question: who got the first nobel prize in physics
Answer:

Desired Answer

Wilhelm Conrad Röntgen



- **LLMs are better at extracting answers from the start and the end (U-shape curve)**
- **Lost in the middle: When the ground truth is in the middle, the accuracy drops**

[5] Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *Transactions of the Association for Computational Linguistics* 12 (2024): 157-173.

Motivation

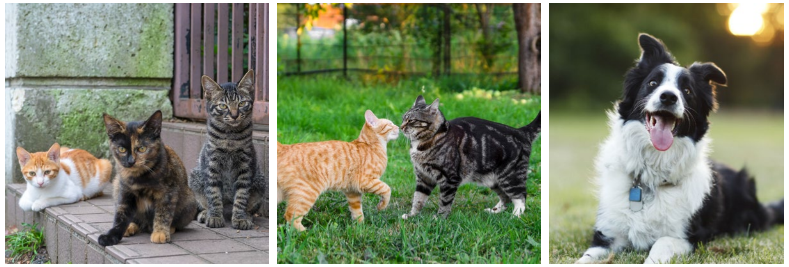
Does such kind of imbalanced reasoning also occur in VLMs?

Question: Can you observe the differences between <image 1> and <image 2>?

🦩 : The first image depicts a night sky and water while the second image features a darker sky with snow covering on the ground.

Question: Can you observe the differences between <image 2> and <image 1>?

🦩 : The two images you provided appear to be identical. There are no observable differences between them.

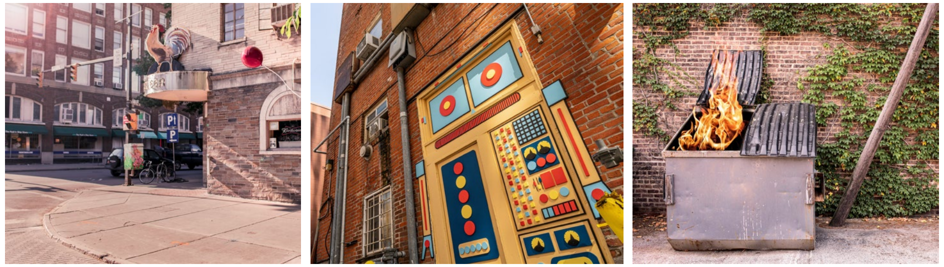


Question: How many cats are there in <image 1>, <image 2> and <image 3> in total?

🤗 : There are three and two cats in the first and second image respectively, while there is only a dog in the third image, therefore five cats in total.

Question: How many cats are there in <image 3>, <image 2> and <image 1> in total?

🤗 : I can see one cat in the first image, two cats in the second, and three cats in the last one. There are totally six cats in these images.



Question: Among <image 1>, <image 2> and <image 3>, which image looks most realistic and not synthesized?

: The first image is most natural. The second and third images depict a mural of painted robot on the building and a dumpster with fire, which displays surrealism and exaggeration.

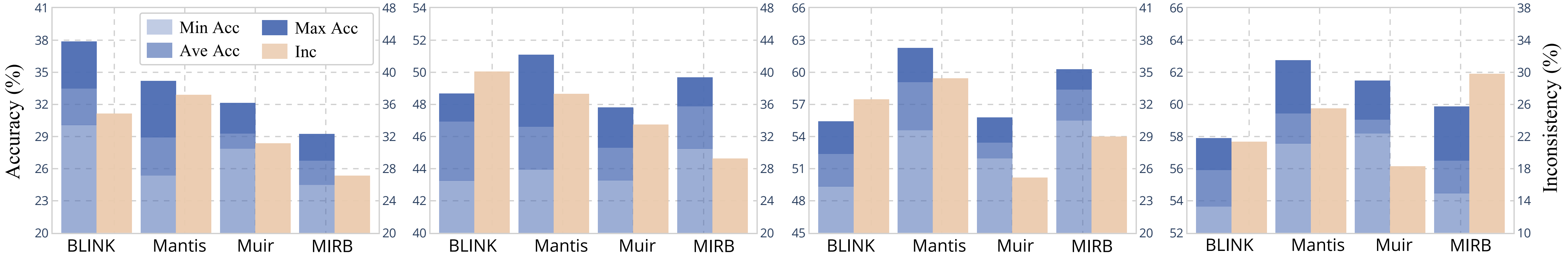
Question: Among <image 3>, <image 1> and <image 2>, which image looks most realistic and not synthesized?

: Overall, the first image looks most realistic since dumpsters can catch fire, while the red ballon and the robot painting in the other two images render a digital and artistic style.

Position bias: VLMs predict the answers significantly relying on the position or order of input image sequences, leading to poor robustness and reliability.

The Position Matters

We select tasks from existing benchmarks whose answers are independent from image positions, then analyze how varying that image order impacts performance.



(A) OpenFlamingo

(B) Idefics2

(C) LLaVA-NeXT

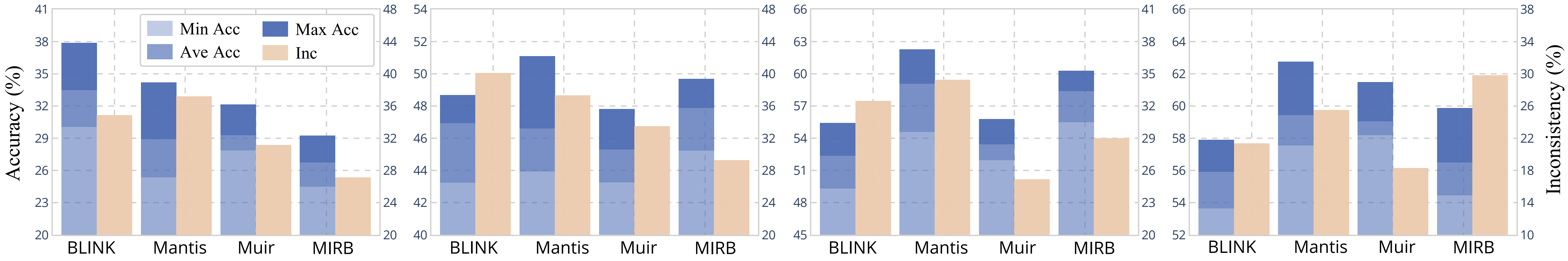
(D) GPT-4o

- **We run multiple evaluations, each with a different random ordering of the images, and report the minimum, maximum, and mean accuracy across these runs.**

- **We also measure prediction inconsistency, defined as the share of examples that receive conflicting answers between the best and worst performing evaluations.**

The Position Matters

We select tasks from existing benchmarks whose answers are independent from image positions, then analyze how varying that image order impacts performance.



(A) OpenFlamingo

(B) Idefics2

(C) LLaVA-NeXT

(D) GPT-4o

- **In most cases, the prediction inconsistency reaches around 30%, indicating one-quarter of examples receive conflicting answers by altering image positions.**

- **As a result, this severe change of predictions lead to an accuracy span of 4~6% in average, which is significant given the overall performance of VLMs.**

Identifying Position Bias

Let’s examine position bias more closely: which image positions do the model handle well, and which ones reveal its weaknesses?

Question: Can you observe the differences between <image 1> and <image 2>?

Question: How many cats are there in <image 1>, <image 2> and <image 3> in total?

Question: Among <image 1>, <image 2> and <image 3>, which image looks most realistic and not synthesized?

However, previous benchmarks only evaluate the holistic understanding of VLMs.

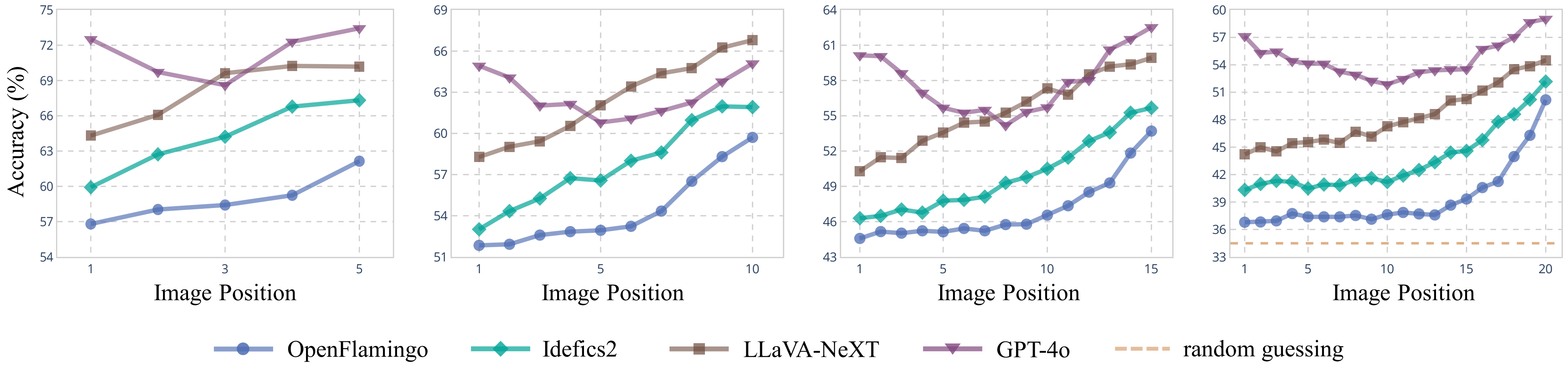
We introduce Position-wise Question Answering (PQA) to evaluate the position-wise reasoning capability of VLMs.

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

- **PQA requires VLMs to produce position-wise response, thereby enabling to track position-wise accuracy.**
- **A higher accuracy indicates stronger reasoning capability given the position, while lower indicates poor-performing areas.**

Identifying Position Bias

We collect 1000 PQA examples from VQAv2, where each PQA example has 5, 10, 15 or 20 images.

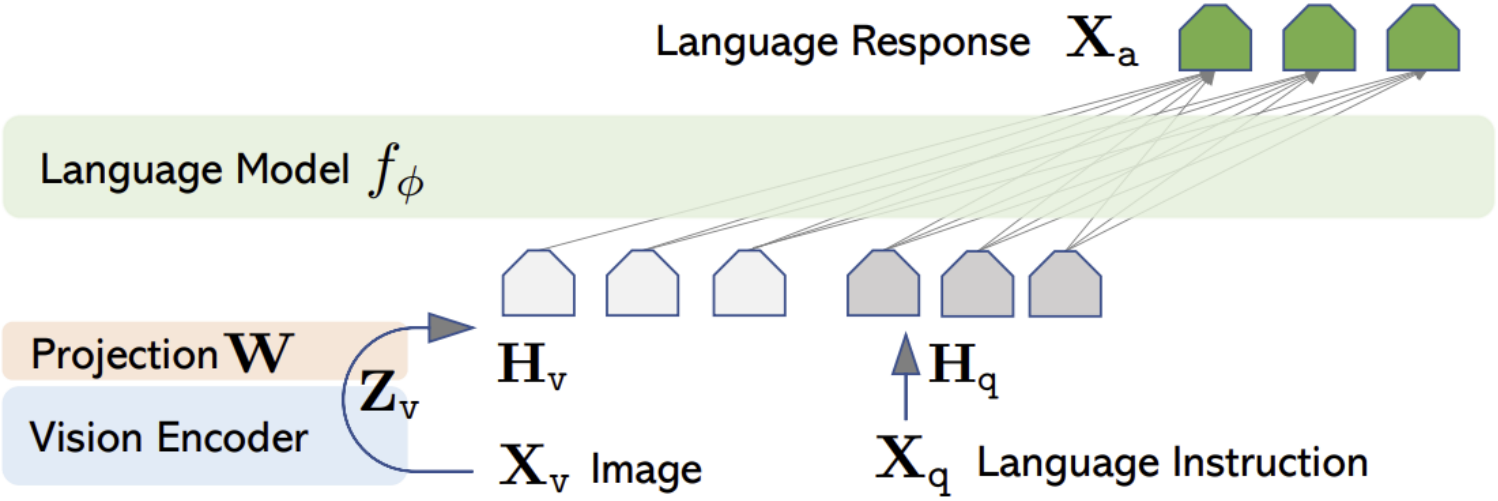


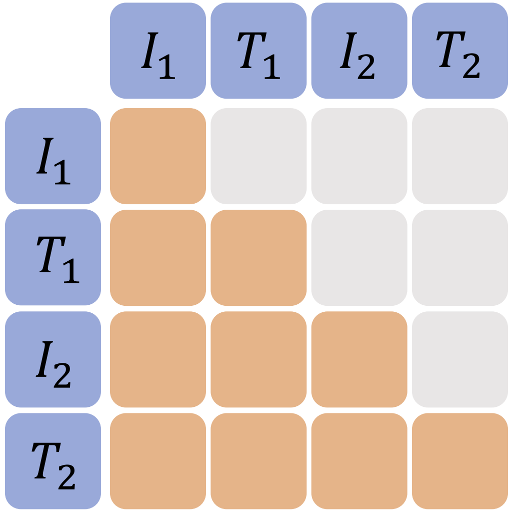
- **For open-source VLMs, we identify recency bias, indicating that VLMs are good at extracting information in the position at the end, while the performance is decaying from back to front.**

- **For proprietary models such as GPT-4o, the conclusion is similar to previous work, i.e., lost in the middle, where the middle part tends to be ignored by VLMs.**

Mitigating Position Bias

Since each VLM have its own featured architecture, we consider a most common type: autoregressive architecture[6].





	I_1	T_1	I_2	T_2
I_1	🟧	⬜	⬜	⬜
T_1	🟧	🟧	⬜	⬜
I_2	🟧	🟧	🟧	⬜
T_2	🟧	🟧	🟧	🟧

Causal Mask

In this case, images are handled like texts by covering a causal mask, which enforces their interaction as a unidirectional information flow:

- **The image at the back may interact with preceding image contexts**
- **The image at the front is isolated, lacking global information**

[6] Laurençon H, Tronchon L, Cord M, et al. What matters when building vision-language models?[J]. Advances in Neural Information Processing Systems, 2024, 37: 87874-87907
[7] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2023, 36: 34892-34916.

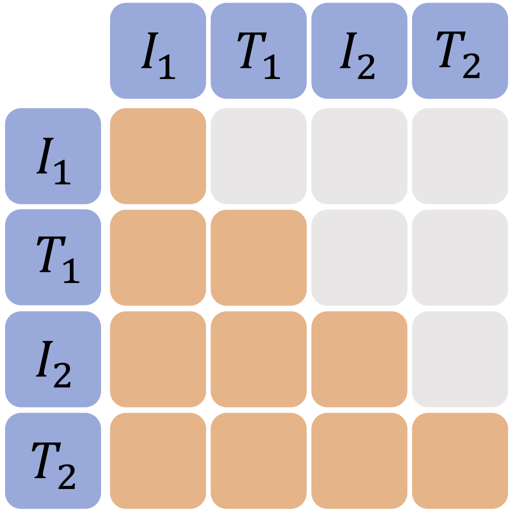
Mitigating Position Bias

The influence of causal masks on image tokens

- **Makes their hidden states position-dependent**
- **Implicitly inject positional information on different images**

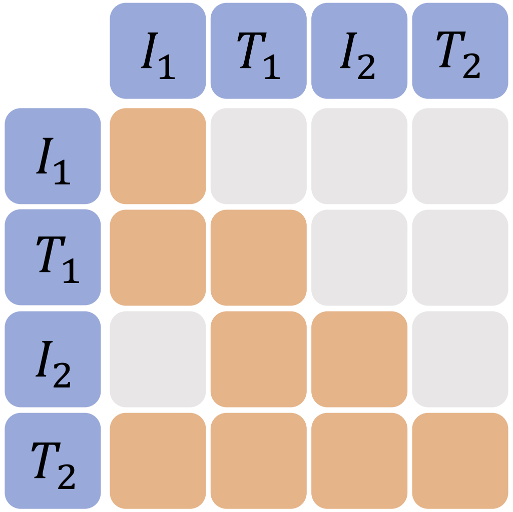
Given the above hypothesis, we compare three variants:

- ***Causal Mask*: Each image can only interact with preceding images.**
- ***Isolated Mask*: Each image can only interact with itself.**
- ***Bidirectional Mask*: Each image can interact with arbitrary images in the sequence.**



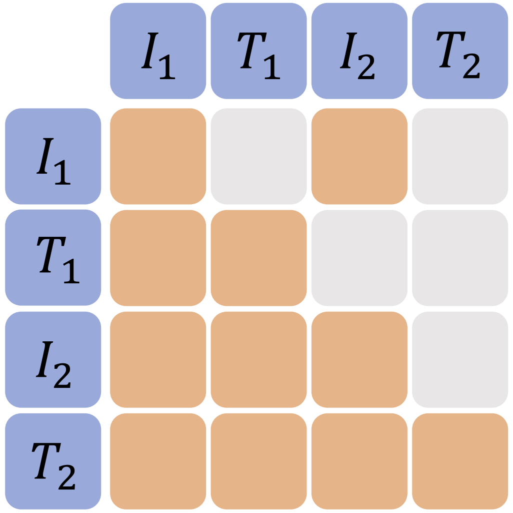
	I_1	T_1	I_2	T_2
I_1	🟧	⬜	⬜	⬜
T_1	🟧	🟧	⬜	⬜
I_2	🟧	🟧	🟧	⬜
T_2	🟧	🟧	🟧	🟧

Causal Mask



	I_1	T_1	I_2	T_2
I_1	🟧	⬜	⬜	⬜
T_1	🟧	🟧	⬜	⬜
I_2	⬜	🟧	🟧	⬜
T_2	🟧	🟧	🟧	🟧

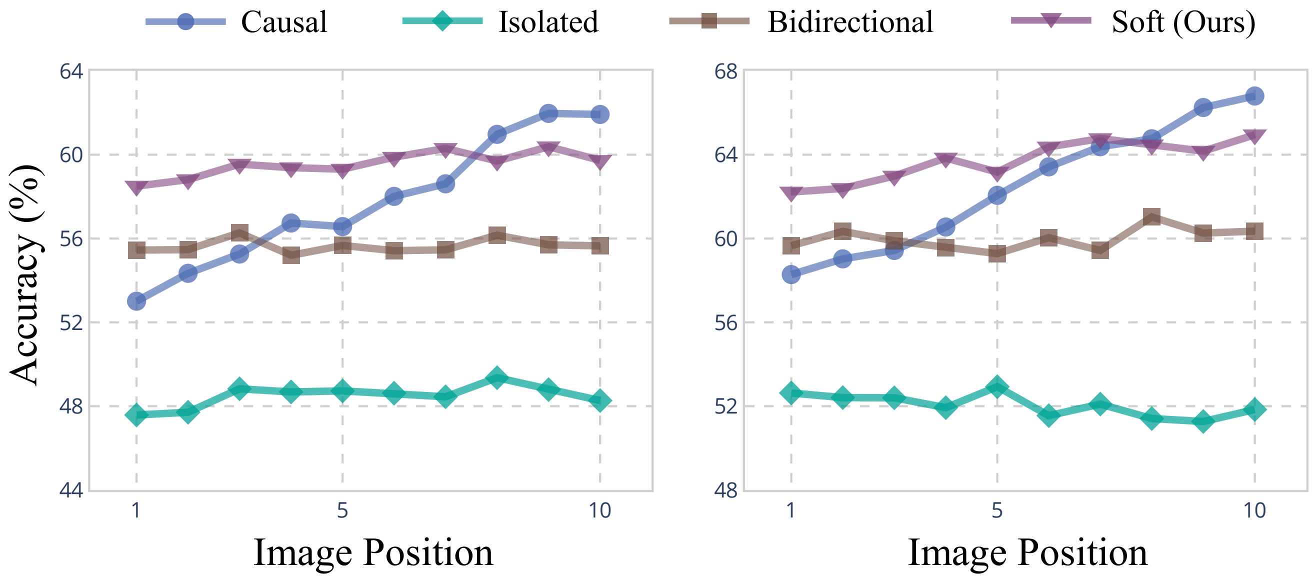
Isolated Mask



	I_1	T_1	I_2	T_2
I_1	🟧	⬜	🟧	⬜
T_1	🟧	🟧	⬜	⬜
I_2	🟧	🟧	🟧	⬜
T_2	🟧	🟧	🟧	🟧

Bidirectional Mask

Mitigating Position Bias



(A) Idefics2 (B) LLaVA-NeXT

- **The unidirectional way of information flow (causal) is the main cause of position bias of VLMs.**

- **Disabling cross attention (isolated) leads to dramatic performance degradation.**

- **The two-way interaction (bidirectional) among images mitigates position bias, at the cost of overall accuracy.**

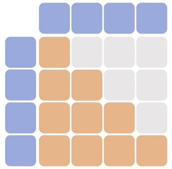
Mitigating Position Bias

The trade-off between accuracy and consistency sparks SoFt Attention (SoFA), a simple method to smoothy position bias by interpolating across attention masks.

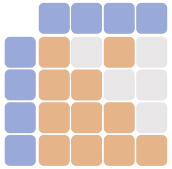
$$H = Softmax\left(\frac{QK^T}{\sqrt{d}}\right) \odot W_{soft} V$$

$$W_{soft} = (1 - \sigma)\, W_{causal} + \sigma\, W_{bidirect}$$

$$W_{causal} =$$



$$W_{bidirect} =$$

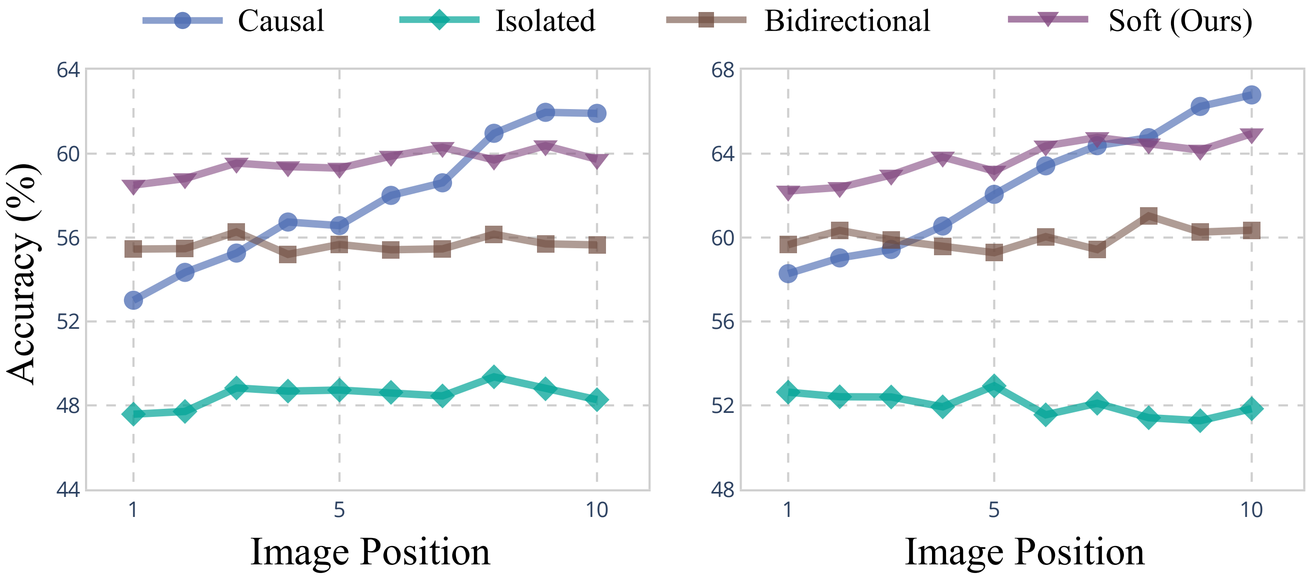


σ controls the strength of mitigation effect:

- **If σ is too small, the position bias cannot be effectively reduced.**
- **If σ is too large, the result mask is out-of-distribution, causing performance drop.**

Mitigating Position Bias

SoFA requires a small validation set per task (32-shot in this case) to find the optimal σ. Besides, we deploy SoFA in language decoders every two layers.

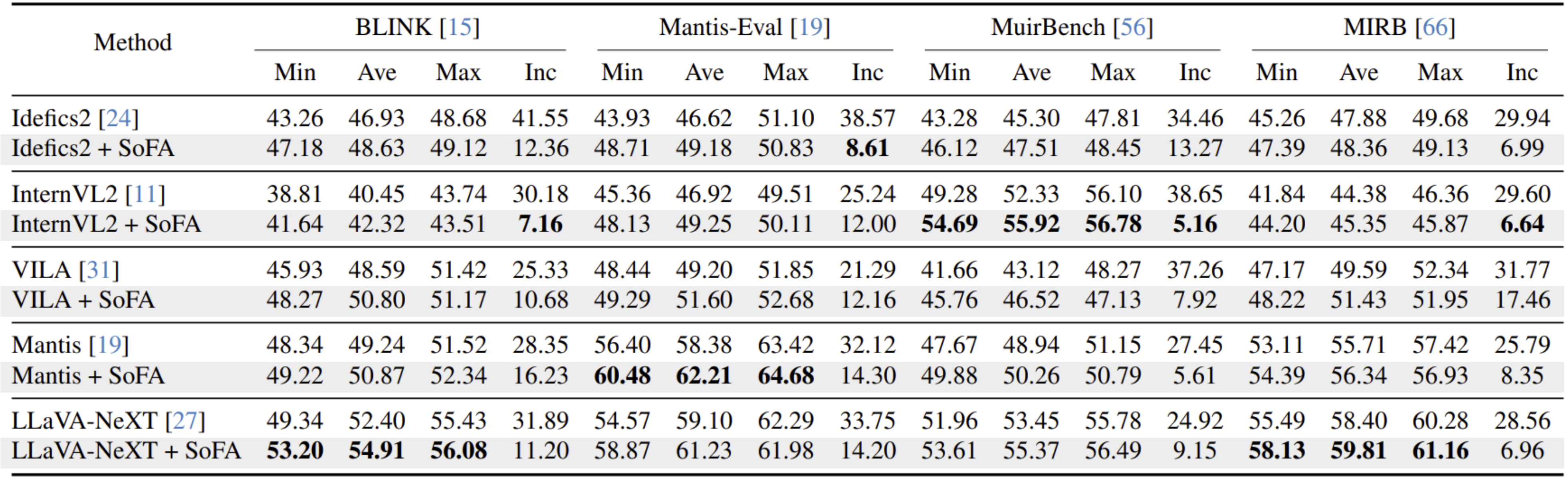


(A) Idefics2

(B) LLaVA-NeXT

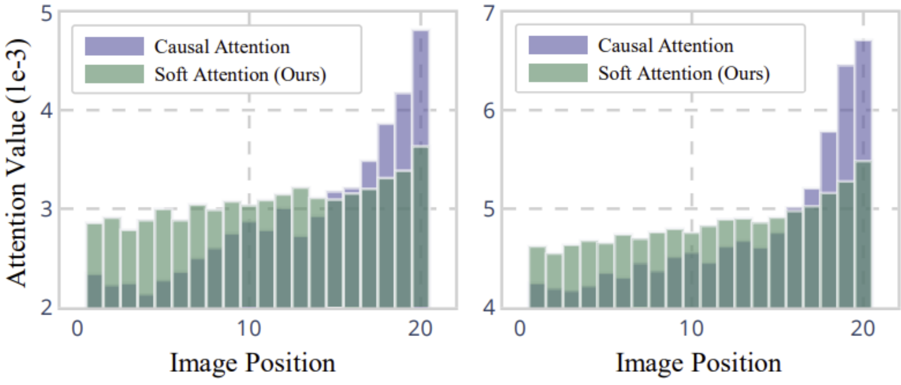
SoFA achieves a satisfying trade-off between performance and bias, reaching a stable and competitive accuracy regardless of image positions.

Experiment & Results



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.



(A) Idefics2 (B) LLaVA-NeXT

Figure 6. The attention distribution across positions on PQA.

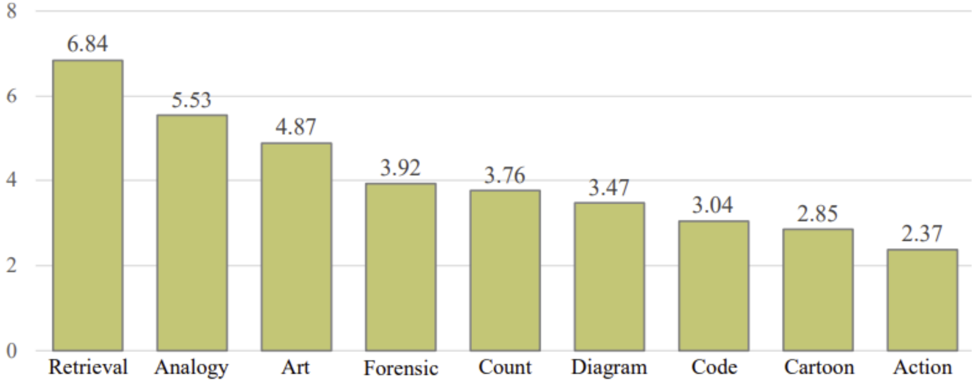


Figure 7. The performance gains of SoFA on different types of tasks. The results are averaged over selected models.

THANK YOU

ExHall D Poster #376

14 June 10:30AM – 12:30PM