Aligning What Vision-Language Models See and Perceive with Adaptive Information Flow

Supplementary Material

Contents

A. Implementation Details

Alg. 1 shows the pseudo-code of our approach. The standard inference process is highlighted in magenta, and information flow modulation is highlighted in blue.

Algorithm 1 Inference with Adaptive Information Flow

```
# I: input image
# T: text prompt
# C: standard causal mask

# get visual and text tokens
v = vision_encoder(I)
t = text_tokenizer(T)

# one-step LLM decoding
A = LLM(v, t, C)   # text-image attention
# entropy of visual tokens
entropy_v = token_entropy_computation(A)
# get indices of masked visual tokens
mask_token_ids = token_masking(entropy_v, A)
# causal mask modulation
# t_s/t_end: start/end index of text tokens
C[t_s:t_end, mask_token_ids] = -∞

# inference with modulated causal mask
generated_text = LLM(v, t, C)
```

Table S1. Breakdown of inference time with LLaVA-1.5-7B on RealWorldQA (in milliseconds). The total inference time includes standard VLM inference (in magenta) and information flow modulation (in blue). The time is measured on a NVIDIA A100 GPU.

Standard Inference L6/7/20	LLM Decoding L10	Token Entropy L12	Token Masking L14/17	Total Time
100ms	30ms	0.15ms	4ms	134.15ms

Inference. Given an input image and text prompt, the vision encoder (including vision-to-text projector) and text tokenizer are adopted to generate visual and text tokens. Based on these tokens, we perform one-step LLM decoding to acquire token dynamics. Next, we compute token entropy using Eq. 4. Subsequently, adaptive token masking is applied to determine what visual tokens should be masked. By modulating the causal mask, the attention between these masked visual tokens and text tokens will be blocked. Finally, LLM will output language responses under the guidance of the modulated causal mask.

For visual token masking, our approach first uniformly samples a set of candidate mask ratios and then selects an appropriate ratio to obtain masked visual tokens. Here we simply sample mask ratios from 0.1 to 0.9 with an interval of 0.1. Note that for token entropy computation and token masking, the implementation is operated at the tensor level, which is efficient.

Computational Cost. Table S1 reports the inference time of our approach with LLaVA-1.5-7B [13] on the RealWorldQA [19] dataset. LLM decoding (Line 10) takes approximately the time of one-token generation. Token entropy computation (Line 12) and token masking (Line 14/17) are relatively computational efficient. Overall, compared with standard VLM inference (L6/7/20), our approach introduces only a marginal computational cost (roughly the time of one-token generation) while significantly improving the performance of baseline.

B. More Results and Discussions

B.1. Discussion on Token Masking

As mentioned in the implementation details, we select an appropriate mask ratio for token masking. Here we provide a deeper analysis of the properties of token masking.

Table S2. Comparisons with baselines on various datasets.

Method	General VQA			OCR Understanding		Counting
	V*	RealWorldQA	MMStar	$\text{TextVQA}_{\text{val}}$	SeedBench2-Plus	CountBench
LLaVA-1.5-7B [13]	42.4	55.6	33.1	47.8	41.3	47.0
LLaVA-1.5-7B AIF (**Ours**)	50.3 +7.9	60.5 +4.9	39.5 +6.4	49.9 +2.1	44.9 +3.6	50.1 +3.1
Qwen2.5-VL-7B [1]	78.5	68.5	63.9	84.9	70.4	87.1
Qwen2.5-VL-7B AIF (**Ours**)	84.8 +6.3	74.5 +6.0	70.9 +7.0	86.0 +1.1	76.5 +6.1	89.5 +2.4

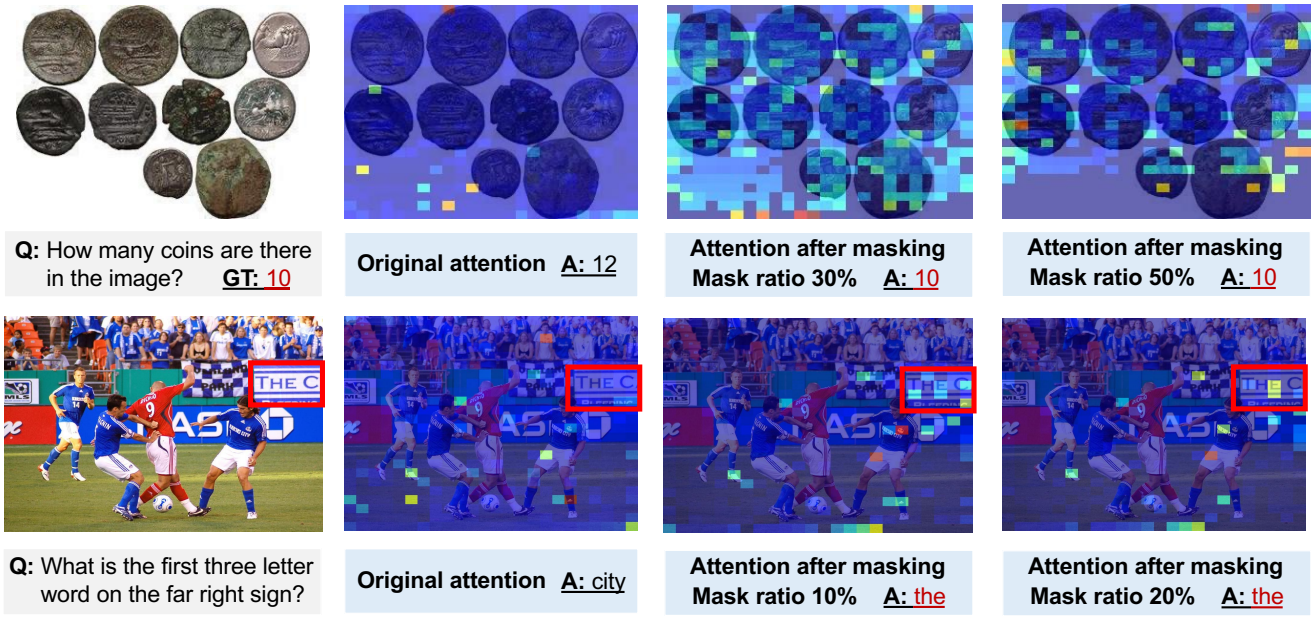


Figure S1. **Results of information flow modulation with different mask ratios.** For an image-question pair, there often exist multiple mask ratios that can improve the answer of the baseline.

Non-Uniqueness of Optimal Mask Ratio. Interestingly, we observe that *there often exist multiple mask ratios that can improve the results of VLMs* for an image. Fig. S1 shows some examples of information flow modulation with different mask ratios. When counting the number of coins, the baseline model incorrectly estimates the count as 12. In contrast, applying token masking at both 30% and 50% ratios enables the model to produce the correct count. A similar improvement is observed in OCR understanding (second row in Fig. S1), where token masking leads to more accurate text recognition. The reason is likely that only a small number of high-attention, distracting visual tokens significantly mislead the model. By removing these irrelevant visual tokens (even in small quantities), the model is able to redirect its attention toward important image regions and produce correct answers. This observation suggests that the selection of optimal mask ratio is flexible.

B.2. The Trend of Improvement

The comparisons with baselines are shown in Table S2. One can observe that: i) datasets with multi-choice questions (*e.g.*, V* [18], RealWorldQA [19], and MMStar [3]) benefit significantly from adaptive information flow. The reason is that the proposed method takes advantage of the information in the question and the options to find important image regions. This helps the model distribute more attention to related objects, resulting in improved answers. ii) the proposed method considerably improves the results on fine-grained tasks such as OCR understanding and counting, but the performance gain is less than that of general VQA.

Overall, information flow modulation provides new opportunities to improve the perception capability of VLMs. On one hand, adaptive information flow consistently improves the results of VLMs across different tasks. On the other hand, given that our approach excels at extracting relevant image regions from the question and options, it could

Table S3. More comparisons with existing methods. All methods use LLaVA-1.5-7B as the base model. The best performance is in **bold** and the second best is underlined.

Method	Re-Training	V*	CountBench	$\text{TextVQA}_{\text{val}}$
LLaVA-1.5	-	42.4	47.0	47.8
CCA [20]	✓	47.1	43.1	45.9
ViCrop [22]	✘	49.7	41.1	**56.1**
Ours	✘	**50.3**	**50.1**	49.9

be beneficial to incorporate possible answers in the question to reduce the difficulty of question answering.

B.3. More Comparisons with Existing Methods

Comparisons with ViCrop and CCA. Table S3 shows that both ViCrop and CCA improve V* [18] but underperform the LLaVA-1.5 baseline on CountBench [15]. We attribute this to how each method reshapes visual evidence. Note that for a fair comparison, we report the performance of ViCrop on the V* dataset without high-resolution visual cropping.

ViCrop [22] enhances local perception by cropping a highly relevant region and feeding the crop together with the original image, which is beneficial when a single local area determines the answer (*e.g.*, TextVQA [17]). However, CountBench requires reasoning over all instances of a target object. Once the model is biased toward the cropped region, objects outside the crop or near the borders are easily ignored, so the model often counts mainly within the selected crop and underestimates the true cardinality. This explains the degradation from 47.0 to 41.1 on CountBench despite gains on TextVQA.

CCA [20] reassigns visual token positions into concentric rings and applies a matching causal mask so that tokens near the image center are relatively closer to the instruction under rotary position Embedding (RoPE). This center-oriented alignment helps reduce hallucination on V* by emphasizing discriminative central regions. At the same time, MCA-LLaVA [24] argues that CCA relies on a heuristic redirection of instruction tokens toward the image center region, without fully resolving RoPE's underlying image-alignment bias. As a result, objects located away from the center can receive systematically weaker attention. In CountBench, where similar objects are spread across the entire scene, this bias degrades the counting capability of VLMs, leading to a drop from 47.0 to 43.1. In contrast, our method preserves the global coverage of LLaVA-1.5 while modulating information flow, and thus achieves the best performance on both V* and CountBench without retraining.

Comparisons with OPERA and VCD. We further compare with two training-free hallucination methods on the POPE [11] dataset. As shown in Table S4, OPERA [6] and VCD [9] achieve accuracies of 85.7 and 86.1, respectively. Their marginal improvements over LLaVA-1.5-7B (85.4) have also been reported in [12]. In contrast, our method achieves an accuracy of 88.7, confirming its effectiveness in mitigating hallucinations.

Table S4. Comparisons with VCD and OPERA on POPE.

LLaVA-1.5	VCD [9]	OPERA [6]	**Ours**
85.4	85.7	86.1	**88.7**

Table S5. Comparisons with irrelevant token pruning, where irrelevant visual tokens identified by our approach are pruned during inference. All methods use LLaVA-1.5-7B as the base model.

Method	RealWorldQA	CountBench
Baseline	55.6	47.0
Irrelevant Token Pruning	$55.4_{-0.2}$	$47.0_{-0.0}$
Ours	$60.5_{+4.9}$	$50.1_{+3.1}$

B.4. Comparison with Token Pruning

Difference with Token Pruning. Token pruning aims to optimize the computational efficiency of VLMs by pruning visual tokens. Existing token pruning methods remove pruned tokens from LLM decoding, resulting in *information loss* and *degraded performance*. In contrast, our goal is to improve the perception capability of VLMs with adaptive information flow. We only disconnect the interaction between text tokens and masked visual tokens, while the remaining visual tokens still have access to these masked visual tokens. This ensures *intact information flow* within visual tokens and leads to *improved performance*.

Pruning vs. Masking Irrelevant Tokens. To quantify the difference between token pruning and information flow modulation, we first identify irrelevant visual tokens using our approach and then prune these token during inference. Table S5 reports the results. One can observe that pruning visual tokens achieves similar performance with the baseline on RealWorldQA [19] and CountBench [15] datasets. The average pruning ratios on RealWorldQA and CountBench are around 30% and 40%, respectively. This suggests that token pruning does not necessarily improve performance. In contrast, our approach consistently boosts the performance of baseline, which highlights the importance and effectiveness of information flow modulation.

B.5. Discussion on Visual Attention Sink

Kang et al. [7] investigate the visual attention sink phenomenon in VLMs. They show that sink tokens (*i.e.*, ir-

relevant visual tokens receiving high attention) can be identified through specific dimensions of the hidden states in LLM. Our work shares a similar background with [7], where VLMs often "see" the relevant objects. However, our method is distinct from it in several aspects: (1) we introduce adaptive information flow that allows VLMs to focus on relevant object regions, whereas [7] first identifies visual sink tokens and then redistributes visual attention; (2) the important tokens selected by our method constitute only a subset of visual tokens, yet they have a significant impact on the accuracy (notably, we observe that high-entropy tokens may correspond to background regions, irrelevant objects, or sink tokens); and (3) for different vision tasks, [7] uses 10% of the test data for hyperparameter tuning, contrary to our method requiring no test data for tuning. This offers new insights over [7] for the community.

B.6. Failure Cases

We notice two types of failure cases: i) when dealing with spatial reasoning (Fig. S2(a)), while our approach could highlight the dog, it could be challenging to reason the relation between the dog and the door. This limitation is consistent with prior findings showing that spatial understanding is one of the most difficult capabilities for VLMs [2, 4, 16, 23]. ii) perception under abnormal conditions is also challenging (Fig. S2(b)). Even if the method successfully highlights relevant text regions, the model may fail to produce correct answers due to its inherent perception capability.

C. Evaluation Details

C.1. Datasets

General Visual Question Answering. We use three datasets, including V* [18], RealWorldQA [19], and MMStar [3].

OCR Understanding. We use TextVQA [17] and SeedBench2-Plus [10]. When evaluating LLaVA-1.5-7B on the TextVQA dataset, we follow ViCrop [22] to directly pass image and question to VLMs, without providing any OCR extracted texts. This setting evaluates the true perception capability of LLaVA-1.5-7B.

Grounding and Counting. We use RefCOCO [8, 14, 21] and CountBench [15].

Hallucination. We adopt POPE [11] dataset.

C.2. Evaluation Prompts of LLaVA-1.5-7B

We use the same prompt on all benchmarks following ViCrop [22]. The prompt format is shown below.

> USER: <image> {question} Answer the question using a single word or phrase. ASSISTANT:

C.3. Evaluation Prompts of Qwen2.5-VL-7B

Unless otherwise specified, we use the default prompt provided by VLMEvalKit [5]. Detailed prompts for all datasets are listed below.

V*

> <image> Question: {question} Options: {options} Answer the question using a single word or phrase.

RealWorldQA

> <image> Question: {question} Options: {options} Please select the correct answer from the options above.

MMStar. We use the following prompt to ensure that the baseline performance of Qwen2.5-VL-7B is consistent with the one reported in [1].

> <image> Question: {question} Options: {options} Answer the question using a single word or phrase.

TextVQA

> <image> {question} Answer the question using a single word or phrase.

SeedBench-2-Plus

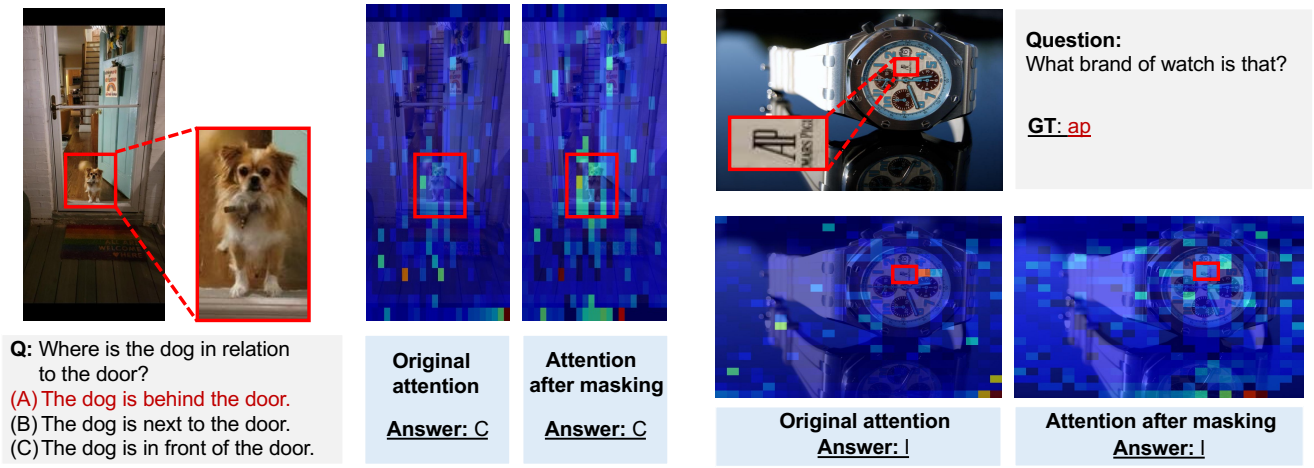
> <image> Question: {question} Options: {options} Please select the correct answer from the options above.

CountBench

> <image> {question} Note that: answer with a number directly e.g. 3. Do not include any additional text.

POPE

> <image> {question} Answer the question using a single word or phrase.



(a) Spatial reasoning **(b) Perception under challenging conditions**

Figure S2. Failure cases of LLaVA-1.5-7B. While our approach is able to highlight object regions to some extent, sometimes it could be challenging to answer correctly when dealing with spatial reasoning and abnormal conditions.

RefCOCO. We follow the official response in a GitHub issue[1] to formulate the prompt as follows:

> <image> Locate {referring_expression} in this image and output the bbox coordinates in JSON format.

D. More Qualitative Results

Here we provide more qualitative results of information flow modulation. As shown in Fig. S3, our approach successfully identifies object regions related to the question, leading to improved answers. This supports the effectiveness of our approach.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2. 5-vl technical report. *arXiv preprint arXiv: 2502.13923*, 2025. 2, 4

[2] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, pages 14455–14465, 2024. 4

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, pages 27056–27087, 2024. 2, 4

[4] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? An attention mechanism perspective on focus areas. In *ICML*, pages 9910–9932, 2025. 4

[5] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACM MM*, pages 11198–11201, 2024. 4

[6] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, pages 13418–13427, 2024. 3

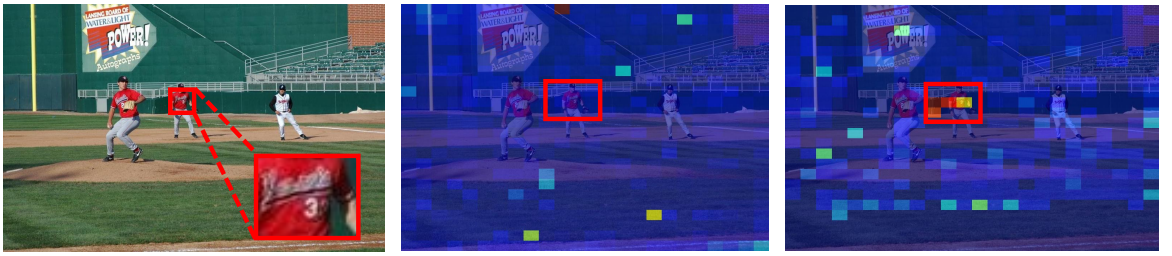
[7] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *ICLR*, 2025. 3, 4

[8] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 4

[9] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pages 13872–13882, 2024. 3

[10] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv: 2404.16790*, 2024. 4

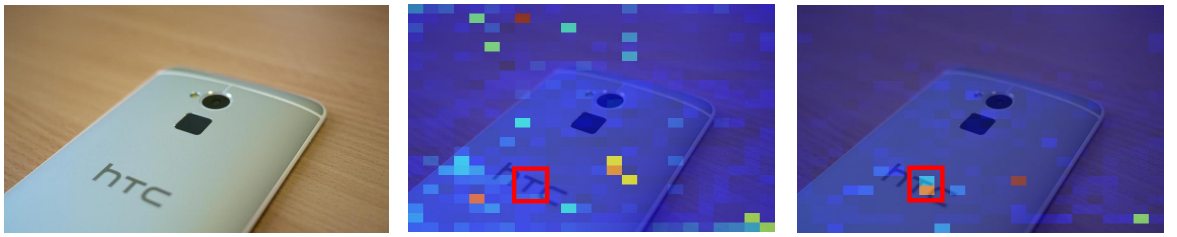
[1] https://github.com/QwenLM/Qwen3-VL/issues/991#issuecomment-2747683955



Q: What is the number of the player in the middle?
GT: 3

Original attention
Answer: 11

Attention after masking
Answer: 3



Q: What letter is in the middle?
GT: t

Original attention
Answer: c

Attention after masking
Answer: t

Q: What is the name of the drink in red?
GT: coca cola

Original attention
Answer: coke

Attention after masking
Answer: coca cola

Q: What is the color of the body of the bucket?
(A) green (B) white
(C) black (D) red
GT: B

Original attention
Answer: D

Attention after masking
Answer: B

Figure S3. More qualitative results of information flow modulation.

[11] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023. 3, 4

[12] Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *ICML*, pages 35799–35819, 2025. 3

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 1, 2

[14] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 4

[15] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *ICCV*, pages 3170–3180, 2023. 3, 4

[16] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *EMNLP*, pages 21440–21455, 2024. 4

[17] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 3, 4

[18] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, pages 13084–13094, 2024. 2, 3, 4

[19] xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. 1, 2, 3, 4

[20] Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. In *NeurIPS*, pages 92012–92035, 2024. 3

[21] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 4

[22] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *ICLR*, 2025. 3, 4

[23] Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *ICLR*, 2025. 4

[24] Qiyan Zhao, Xiaofeng Zhang, Yiheng Li, Yun Xing, Xiaosong Yuan, Feilong Tang, Sinan Fan, Xuhang Chen, Da-Han Wang, and Xu-Yao Zhang. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. In *ACM MM*, page 3981–3990, 2025. 3