

Cross-Modal Attention Calibration for LVLMM Hallucination Mitigation (Supplementary Materials)

1. More Comparisons with Existing Methods

Detailed Comparison on POPE under VCD Settings. We report the detailed performance on POPE in Table 1 with nucleus sampling under VCD settings for further comparison. As presented in the table, CMAC leads to performance improvement on both precision and recall on most of setups, which demonstrates the generalization capability of CMAC to generate both precise and comprehensive responses.

GPT-4o Assisted Evaluation. To assess the effectiveness of our method in mitigating hallucination beyond object existence during long-sequence generation, we conducted open evaluations of LVLMM-generated responses using GPT-4o, as presented in Table 2. Following PAI [3], we sampled 50 images from the COCO dataset for image captioning tasks. Unlike PAI, which employs GPT4-V, we utilized the recently released GPT-4o to evaluate generated responses in terms of accuracy and detail. The GPT-4o prompts were designed in alignment with the structure in PAI [3]. We compare our method with the baseline method and existing decoding methods, including VCD [2] and ICD [4]. Our proposed method achieved superior GPT-4o scores compared to baseline and existing decoding approaches, reflecting its robustness under a more comprehensive evaluation framework. These results highlight the enhanced capability of our approach to generate more accurate and detailed responses for image captioning tasks.

Evaluation with LLaVA-OneVision. To further assess the generalization capability of our method, we apply our method to the latest LLaVA-OneVision (llava-onevision-qwen2-7b-ov) on the POPE dataset. The results are reported in Table 3. It shows that our method outperforms VCD and consistently enhances the performance of LLaVA-OneVision, underscoring the generalization of our method.

Evaluation on MMBench. We also test the results on the general ability benchmark MMBench on LLaVA-1.5. As in Table 4, our method achieves a clear improvement on the general ability of LVLMMs.

Detailed Comparison on CHAIR. The *max new token* parameter plays a critical role in the evaluation of the CHAIR metric by restricting the maximum length of generated responses. In the main text, we present results under the set-

tings *max new token* = 1024 and *top p* = 1. Additionally, following [1, 5], we also report results for *max new token* = 64 in Table 5. As illustrated in the table, the *max new token* value significantly influences the performance of LVLMMs on the CHAIR metric. Nevertheless, our method outperforms other methods by a clear margin. It shows our method consistently demonstrates superior hallucination mitigation capabilities, further validating its robustness and effectiveness under a strict constraint on the length of generated responses.

Inference Time Analysis. To validate the effectiveness of our method on the inference speed, we estimate the inference time consumption of MLLMs employing different decoding strategies in the POPE dataset in Table 6. The results demonstrate that our proposed method achieves faster inference speeds compared to existing methods such as VCD and ICD. This improvement stems from CMAC’s ability to directly derive the attention weights and key-value vectors for the distorted forward process from the original forward process, avoiding redundant computations.

2. More Ablation and Analysis

Analysis of ROPE with CMPC. To illustrate the advantages of our proposed CMPC, we first analyze the mechanism and limitations of Rotary Position Embedding (RoPE). RoPE encodes positional information by rotating query and key vectors according to their absolute positions. As shown in Equation 1, the rotation for a token at position j and another at position i is applied independently.

$$\mathbf{R}_k^i = \begin{pmatrix} \cos(i\theta_k) & -\sin(i\theta_k) \\ \sin(i\theta_k) & \cos(i\theta_k) \end{pmatrix}, \quad \theta_k = \frac{1}{b^{2k/d}},$$
$$\mathbf{a}_{j,i}^l \propto (\mathbf{q}^j \mathbf{R}^j)(\mathbf{k}^i \mathbf{R}^i)^T = \mathbf{q}^j \mathbf{R}^j (\mathbf{R}^i)^T (\mathbf{k}^i)^T$$
$$= \mathbf{q}^j (\mathbf{R}^{j-i})^T (\mathbf{k}^i)^T, \quad (1)$$

A crucial property of RoPE, derived in Equation 1, is that the final attention score only depends on the relative distance ($j - i$). However, this elegant design faces a critical failure mode when processing sequences longer than the model’s training context. The rotation angle $(j - i)\theta_k$, is proportional to this distance. In high-frequency dimen-

Dataset	Setup	Method	LLaVA 1.5				InstructBLIP			
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
MS-COCO	Random	Baseline	83.29	92.13	72.80	81.33	80.71	81.67	79.19	80.41
		VCD	87.73	91.42	83.28	87.16	84.53	88.55	79.32	83.68
		ICD	85.17	95.99	73.40	83.19	<u>86.43</u>	92.01	80.73	<u>85.61</u>
		Ours	89.23	93.50	84.33	88.68	86.96	93.97	79.00	85.83
	Popular	Baseline	81.88	88.93	72.80	80.06	78.22	77.87	78.85	78.36
		VCD	85.38	86.92	83.28	85.06	81.47	82.89	79.32	81.07
		ICD	82.33	82.25	<u>82.47</u>	82.36	82.93	84.45	80.73	<u>82.55</u>
		Ours	86.90	92.28	80.53	86.01	84.13	88.38	78.60	83.20
	Adversarial	Baseline	78.96	83.06	72.75	77.57	75.84	74.30	79.03	76.59
		VCD	80.88	79.45	83.29	81.33	79.56	79.67	79.39	79.52
		ICD	81.17	87.13	73.13	79.52	80.87	80.95	80.73	80.84
		Ours	83.23	85.28	80.33	82.73	82.57	86.10	77.67	81.67
A-OKVQA	Random	Baseline	83.45	87.24	78.36	82.56	80.91	77.97	86.16	81.86
		VCD	86.15	85.18	87.53	86.34	84.11	82.21	87.05	84.56
		ICD	86.20	91.07	80.27	85.33	<u>85.82</u>	83.80	88.94	<u>86.29</u>
		Ours	88.53	<u>90.19</u>	86.47	88.29	87.03	88.71	84.87	86.75
	Popular	Baseline	79.90	80.85	78.36	79.59	76.19	72.16	85.28	78.17
		VCD	81.85	78.60	87.53	82.82	79.78	76.00	87.05	81.15
		ICD	82.63	84.25	80.27	82.21	<u>81.64</u>	78.50	88.77	83.32
		Ours	85.73	86.66	84.47	85.55	82.60	80.64	85.80	<u>83.14</u>
	Adversarial	Baseline	74.04	72.08	78.49	75.15	70.71	65.91	85.83	75.56
		VCD	74.97	70.01	87.36	77.73	<u>74.33</u>	<u>69.46</u>	86.87	77.19
		ICD	77.20	75.47	80.60	77.95	<u>74.42</u>	<u>70.24</u>	88.93	78.48
		Ours	79.13	76.45	84.20	80.14	75.70	71.38	85.80	<u>77.93</u>
GQA	Random	Baseline	83.73	87.16	79.12	82.95	79.65	77.14	84.29	80.56
		VCD	86.65	84.85	<u>89.24</u>	86.99	83.69	81.84	86.61	84.16
		ICD	85.73	90.79	79.53	84.79	85.10	84.21	86.40	85.29
		Ours	89.53	<u>90.62</u>	88.20	88.80	86.03	87.67	83.87	85.72
	Popular	Baseline	78.17	77.64	79.12	78.37	73.87	69.63	84.69	76.42
		VCD	80.73	76.26	89.24	82.24	78.57	74.62	86.61	80.17
		ICD	79.87	80.07	79.53	79.80	<u>78.80</u>	75.15	87.53	80.87
		Ours	85.50	85.24	85.87	85.55	79.87	77.28	84.60	<u>80.78</u>
	Adversarial	Baseline	75.08	73.19	79.16	76.06	70.56	66.12	84.33	74.12
		VCD	76.09	70.83	88.75	78.78	75.08	70.59	85.99	<u>77.53</u>
		ICD	77.60	76.04	80.60	78.25	75.17	70.59	86.27	<u>77.65</u>
		Ours	81.87	79.18	86.47	82.66	76.83	73.94	82.87	78.15

Table 1. **Results on discrimination hallucination benchmark POPE.** The Baseline method denotes the standard decoding. The best performances within each setting are **bolded**. Comparable (± 1.0) but not the best performances are underlined. We implemented ICD for LLaVA 1.5 in our evaluation setup.

Method	Accuracy \uparrow	Detailedness \uparrow
Baseline	5.38	5.88
Ours	6.65	6.45
VCD	5.83	5.93
Ours	6.57	6.54
ICD	6.03	6.48
Ours	6.61	6.52

Table 2. The results of LLaVA 1.5 for GPT-4o assisted evaluation. All the metrics are on a scale of 1 to 10.

sions (i.e., for small k , where θ_k is large), a large relative distance can cause the angle to exceed 2π .

Once the angle surpasses this π threshold, the trigono-

metric functions are no longer monotonic, leading to ambiguity. For instance, the model cannot distinguish a relative distance from another distance if their corresponding rotation angles map to the same cosine value, as shown in Figure 1. This phenomenon, known as position aliasing, causes the model to assign erroneous attention weights to distant tokens, resulting in a catastrophic degradation of performance.

Our proposed CMPC alleviates this issue. By systematically scaling the position indices of all tokens, CMPC effectively compresses the large rotation angles to low angles that the model can correctly interpret. This enhances the model’s representation capability by preventing position aliasing and preserving its ability to accurately model long-range dependencies.

Method		Sampling						Greedy					
		Random		Popular		Adversarial		Random		Popular		Adversarial	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LLaVA-OV	Baseline	87.18	86.07	84.07	83.40	80.99	80.90	90.10	89.30	87.81	87.17	84.17	84.02
	VCD	88.69	87.78	86.18	85.56	82.89	82.38	90.28	89.44	88.57	87.83	84.20	83.86
	Ours	89.52	88.83	87.32	86.71	83.69	83.40	90.45	89.50	89.12	88.38	84.93	84.67

Table 3. Results on discrimination hallucination benchmark POPE. The Baseline method denotes the standard decoding. The best performances within each setting are **bolded**. "Acc" and "F1" denote the Accuracy and F1 scores, respectively.

Method	MMBench Score
Baseline	63.9
Ours	64.7

Table 4. Results of LLaVA1.5 on general ability benchmark MM-bench.

Method	CHAIR _i ↓	CHAIR _s ↓	Recall ↑	F1 Score ↑
Baseline	25.4	9.1	56.7	69.8
VCD	22.0	6.9	62.0	74.4
ICD	22.2	7.8	61.1	73.5
Ours	20.3	5.9	61.9	74.7

Table 5. Results of CHAIR on the MS-COCO validation set with the *max new token* = 64.

Method	Inference speed (tokens/s)
Baseline	12.93
VCD	7.04
ICD	6.90
Ours	7.75

Table 6. Ablation of inference speed in our method.

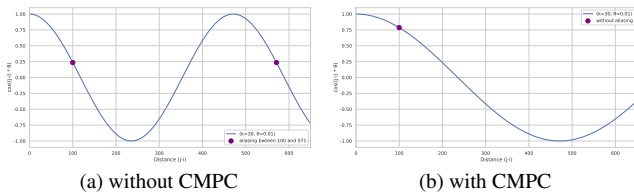


Figure 1. The rotation value of ROPE under different token distances without (a) and with (b) the CMPC.

Analysis of ROPE with CMPC. To further validate the effectiveness of CMPC, we analyze the attention distribution from text tokens to image tokens with and without position embeddings by visualizing the attention maps. As depicted in Figure 2(a), when position embeddings are included, the text tokens predominantly focus on the latter portion of the visual content, neglecting other relevant regions. It shows that the position embeddings lead to position bias in the attention mechanism. Conversely, as in Figure 2(b), CMPC significantly enhances attention weights to image tokens. We also visualize the attention maps of each head in Fig-

ure 3. It can also be observed that the CMPC helps to encourage LLMs to focus on the crucial information in images. These demonstrate that the CMPC effectively eliminates the position relation among different image tokens on the cross-modal part to mitigate the overlooking of critical visual information.

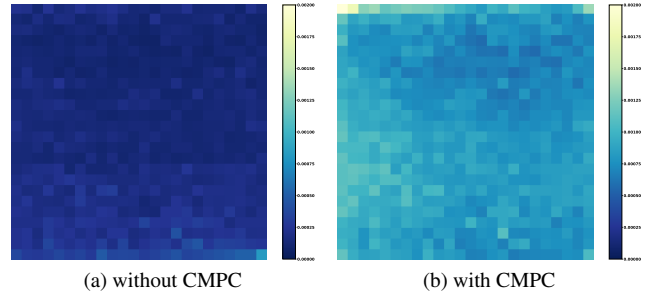


Figure 2. The attention distribution of text tokens to image tokens without (a) and with (b) the position embeddings.

Ablation of variants of CMPC. We conduct more ablation about the variants of the position embedding strategy in CMPC in Table 7. Note that "GPC" demotes global position calibration, which directly scales up all the position indexes on all the attention weights. We find that GPC performs even better than our method on the POPE dataset, indicating the position bias in uni-modal knowledge exchange. However, GPC scales the position indexes for text tokens, which makes the LLMs hard to distinguish the length of generated text and induces more hallucinations. This issue can be alleviated by directly enhancing the probability of EOS tokens by a fixed value to normalize the length of the generated output. GPC + EOS enhancement leads to significant improvement on CHAIR_i but slightly reduces the CHAIR_s. We consider it a variant of our method. Additionally, the refined position embeddings used in CMPC further lead to an improvement on both CHAIR_i and CHAIR_s, which shows the significance of the global positional relation of the image content with other language tokens.

3. Choices of Hyper-parameters

In this section, we analyze the choices of hyperparameters used in the proposed approach, including α , and γ .

Choice of α . α is the most important parameter for the contrastive decoding, which moderates the contrastive effect.

Method	POPE		CHAIR				
	Accuracy	F1 Score	CHAIR _i	CHAIR _s	Recall ↑	F1 ↑	Len
Without CMPC	85.10	84.68	48.2	13.6	75.8	80.7	101.0
GPC	86.57	87.12	61.6	17.1	81.8	82.3	146.0
GPC + EOS enhancement	86.57	87.12	44.8	13.8	75.7	80.6	102.3
CMPC	86.04	86.70	47.0	12.7	75.6	81.0	102.4

Table 7. Ablation of different variants of CMPC module in our method. 'GPC' denotes global position calibration, which directly scales up all the position indexes on all the attention weights. 'EOS enhancement' denotes directly enhancing the probability of the EOS tokens.

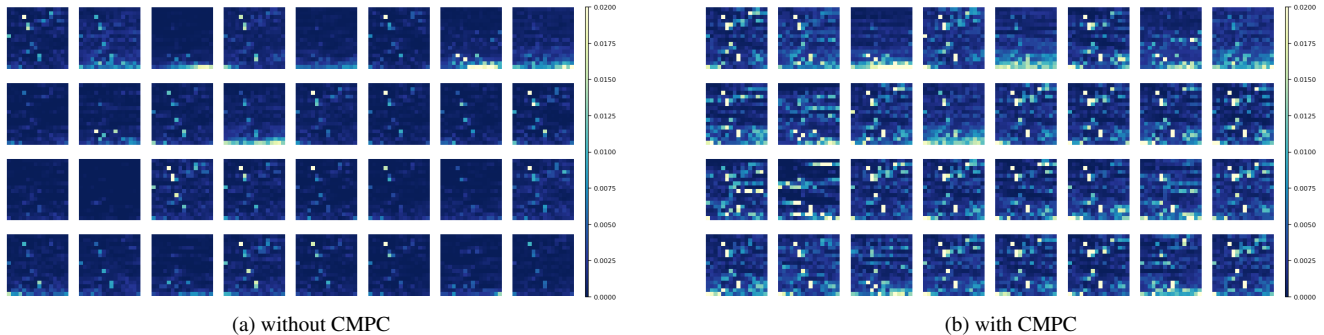


Figure 3. The attention of different heads of text tokens to image tokens with (a) and without (b) the position embeddings.

To determine the value of α , we compared model performance under different α for our method in Table 8. Compared with the performance under $\alpha = 1$, our method shows a better performance under a more aggressive setting $\alpha = 3$ and $\alpha = 5$ with nucleus sampling. However, $\alpha = 5$ leads to a performance degradation in greedy search (85.85 vs. 86.33 at $\alpha = 3$). Consequently, we selected $\alpha = 3$ as the default setting to ensure robust performance across different decoding strategies.

α	Sampling		Greedy search	
	Accuracy ↑	F1 Score ↑	Accuracy ↑	F1 Score ↑
1	85.50	85.12	86.27	85.89
3	86.04	85.70	86.33	86.00
5	86.42	86.33	85.85	85.52

Table 8. The performance of our method with the different values of α on LLaVA 1.5 under POPE popular.

Choice of γ . To better select γ , we present its performance under different settings in Table 9. The results indicate that a larger γ may lead to better performance on POPE. However, the performance on CHAIR is more sensitive to the increase of γ , resulting in a decline in performance. Setting γ to 2 as adopted in our method achieves a better tradeoff between different tasks.

4. Long Sequence Response Examples

To further validate the effectiveness of our method, we present some cases of long sequence responses in image captioning tasks in Figure 4 and Figure 5. These examples illustrate that the proposed CMAC approach effectively mitigates hallucinations in the generated responses, providing

γ	POPE		CHAIR	
	Accuracy ↑	F1 Score ↑	CHAIR _i ↓	CHAIR _s ↓
1.5	85.73	85.66	47.5	12.9
2	86.04	85.70	47.0	12.7
3	86.32	85.88	48.2	13.9
5	86.25	85.91	49.1	14.3

Table 9. The performance of our method under the different values of γ on LLaVA 1.5.

more accurate and contextually relevant descriptions compared to baseline methods.

5. Limitations

Our method notably enhances the inference performance of LVLMs by effectively addressing hallucination issues. In the IMD module, we mitigate spurious inter-modality correlations by selectively masking value vectors based on the magnitude of attention weights. However, the magnitude of attention weights may not fully capture the causal relevance between text and visual tokens. Developing a more refined selection mechanism could enable a more accurate estimation of distorted distributions, further improving performance. Additionally, since CMAC does not enhance the visual encoder's ability to extract and represent relevant knowledge from images, its effectiveness remains constrained by the grounding capability of the visual encoder. Future work will explore these directions to achieve more robust and comprehensive improvements.



Please describe this image in detail.



The image captures a large, majestic elephant walking through a field with a lush green forest surrounding it. The elephant strolls casually along an area of dirt and **is accompanied by a small or juvenile elephant close by, a little smaller than the main one.** The elephants are walking through a flowing river, their trunks up near them, possibly as an indication of their presence or curiosity about the water. The scene constitutes a serene and picturesque snapshot of an elephant-filled forest in its natural habitat.



Normal

The image features an elephant walking across a lush green field, surrounded by a forest. The elephant is positioned towards the right side of the scene, making its way across the landscape. In the background, there is **a bird perched on a tree branch,** adding a touch of wildlife diversity to the scene. The field is filled with grass and appears to be a part of the elephant's natural habitat.



VCD

The image features a large elephant standing alone in a grassy field, close to a watering hole. The elephant is facing the camera and appears to be enjoying the natural surroundings. The scene is lush and green, with trees surrounding the area and a body of water nearby. The elephant's presence in this serene environment highlights the beauty of nature and the majesty of these majestic animals.



Ours



Please describe this image in detail.



A young girl and a boy are standing in their living room playing a video game on a Wii gaming console. The girl appears to be holding a Wii remote, while both children enjoy the game being displayed on the television. The TV is centered towards the background of the scene, and a couch can be spotted behind them, slightly to the right. The room also has multiple books scattered around, likely for the family to read leisurely, showcasing the comfortable and entertaining atmosphere in the living space.



Normal

The image features a living room with a television situated in the middle of the scene. Two children, a boy and a girl, are playing a video game on the Wii console, each holding a remote in their hands. They are standing close to the television, fully engaged in the game. The room also contains several books scattered around, some of which are placed on a bookshelf and others are lying on the floor. **A clock is mounted on the wall, and there is a handbag placed near the right edge of the image. Additionally, there is a bowl and a cup placed in the room, possibly for refreshments during their gaming session.**



VCD

The image depicts a young girl and a boy playing a baseball video game together on a flat screen TV in a living room. The girl is standing on the right side of the scene, while the boy is positioned on the left side. They both hold remotes in their hands, actively engaged in the game. The living room features a TV mounted on the wall, a bookcase filled with various books, and a small basket on a surface near the TV.



Ours

Figure 4. Qualitative results of different methods on MS-COCO image captioning task. The hallucinated content is highlighted in red.



Please describe this image in detail.



In the image, a large bird is perched on top of a parking meter on the side of a busy street. The street is filled with traffic, as various cars, trucks, a bus, and a motorcycle occupy the scene. Some vehicles are parked along the street, while others are in motion. The setting seems to be urban, with traffic lights present in many different parts of the scene. Some city benches are visible, likely sitting near the curb for people to rest or wait. The bird dominates the foreground, standing out from the bustling street environment.



Normal

The image features a crowded street scene with a black bird perched on a parking meter, possibly waiting for its turn to cross the road. The street is filled with various vehicles, including several cars, a bus, and a truck. In addition to the cars, there are two traffic lights visible in the scene, one located towards the left side and another closer to the right. A person can be seen standing near the center of the scene, possibly waiting to cross the street or simply observing the bustling environment.



VCD

The image features a black bird perched on a parking meter next to a car. The bird appears to be looking inside the car window, possibly searching for food. The car is parked on the side of the street, and there are several other vehicles in the scene. In addition to the car with the bird, there are four other cars parked or driving nearby. There are also two traffic lights visible, one at the top left corner and another closer to the center of the scene. A truck can be seen on the left side of the image, possibly indicating ongoing construction or maintenance work in the area.



Ours



Please describe this image in detail.



The scene features a woman standing in the grass in front of a building, throwing something possibly a frisbee or kite for her cat to chase or play with. A cat in the area seems to be watching the kite or toy in motion. Another person can be seen nearby, standing at a distance from the cat. A fire hydrant is visible in the background, located near the right side of the image, and the playground can be seen in the far right corner. The overall setting is an outdoor environment with grassy areas and buildings.



Normal

The image features a young boy flying a kite in a backyard, with a cat watching nearby. The boy is holding onto the kite string, which is visible in the scene. The cat is lying in the grass, observing the kite's movements. In the background, there are two parked cars, one on the left side and the other on the right side of the scene. Additionally, there is a bottle located near the center of the image, possibly belonging to the boy or someone else in the yard.



VCD

In the image, a person is flying a kite in a green yard near a white house. The kite can be seen soaring in the sky, while the person stands on the grass holding the kite string. There is also a cat lying on the grass, watching the kite as it flies overhead. The yard is well-maintained with a sidewalk visible, and a street can be seen nearby. The scene conveys a leisurely and relaxed atmosphere, as the person enjoys the outdoors with their cat.



Ours

Figure 5. Qualitative results of different methods on MS-COCO image captioning task. The hallucinated content is highlighted in red.

References

- [1] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. [1](#)
- [2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. [1](#)
- [3] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024. [1](#)
- [4] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Bie-mann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024. [1](#)
- [5] Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*, 2024. [1](#)