CLAIM: Mitigating Multilingual Object Hallucination in Large Vision-Language Models with Cross-Lingual Attention Intervention

Zekai Ye[1*], **Qiming Li**[1*], **Xiaocheng Feng**[1,2], **Libo Qin**[3], **Yichong Huang**[1], **Baohang Li**[1], **Kui Jiang**[1], **Yang Xiang**[2], **Zhirui Zhang**[4], **Yunfei Lu**[4], **Duyu Tang**[4], **Dandan Tu**[4], **Bing Qin**[1,2]

[1]Harbin Institute of Technology [2]Peng Cheng Laboratory
[3]Central South University [4]Huawei Technologies Co., Ltd
{zkye,qmli,xcfeng}@ir.hit.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have demonstrated impressive multimodal abilities but remain prone to multilingual object hallucination, with a higher likelihood of generating responses inconsistent with the visual input when utilizing queries in non-English languages compared to English. Most existing approaches to address these rely on pretraining or fine-tuning, which are resource-intensive. In this paper, inspired by observing the disparities in cross-modal attention patterns across languages, we propose **Cross-Lingual Attention Intervention for Mitigating Multilingual Object Hallucination (CLAIM)** in LVLMs, a novel near training-free method by aligning attention patterns. CLAIM first identifies language-specific cross-modal attention heads, then estimates language shift vectors from English to the target language, and finally intervenes in the attention outputs during inference to facilitate cross-lingual visual perception capability alignment. Extensive experiments demonstrate that CLAIM achieves an average improvement of 13.56% (up to 30% in Spanish) on the POPE and 21.75% on the hallucination subsets of the MME benchmark across various languages. Further analysis reveals that multilingual attention divergence is most prominent in intermediate layers, highlighting their critical role in multilingual scenarios.

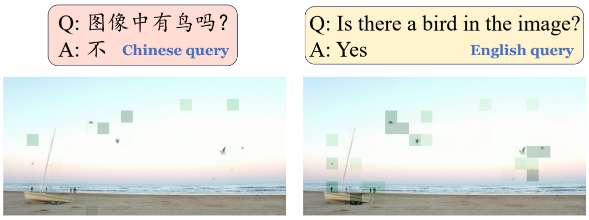


Figure 1: A comparison of attention weights map between Chinese and English query. In English query, LVLM correctly focuses on the key object "bird" in the image, leading to an accurate response. However, in Chinese query, the model exhibits hallucination.

1 Introduction

Large Vision-Language Models (LVLMs) have made significant strides in bridging visual and textual content (Bai et al., 2023b; Liu et al., 2024c; Ye et al., 2023), leading to notable developments in numerous downstream tasks (Shah et al., 2023; Zhu et al., 2023; Zhang et al., 2024). However, LVLMs still suffer from serious object hallucination, *i.e.*, generating responses that are inconsistent with the visual input (S. et al., 2023; Z. et al., 2024; Huang et al., 2023), such as misidentifying the presence of objects in an image or providing inaccurate descriptions of their attributes. This issue becomes even more severe when processing non-English queries (Schneider and Sitaram, 2024; Qu et al., 2024; Romero et al., 2024), a challenge referred to as **multilingual object hallucination** in LVLMs.

Rencent research (Qu et al., 2024) focus on mitigating multilingual object hallucination in LVLMs via adopting Supervised Fine-Tuning (SFT) (Liu et al., 2023) and Direct Preference Optimization (DPO) (Zhao et al., 2023). However, these technologies rely on large-scale annotated image-text datasets, which are extremely expensive, time-consuming, and computation-consuming for non-English especially low-resource languages. Therefore, it is urgent to develop a training-free method for mitigating multilingual object hallucination, with further understanding of the behavioral disparities of LVLMs in multilingual scenarios.

Inspired by prior research (Liu et al., 2025; Bi et al., 2024; Chen et al., 2024b; Jiang et al., 2024) highlighting the crucial role of attention in bridging textual and visual information, we discover the significant difference in attention patterns of LVLMs across languages. Specifically, as illustrated in Figure 1, the model pays attention to distinct areas of the image when processing queries in different

[*] Equal Contribution

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13080–13094
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

languages. Under non-English queries, LVLMs' intermediate layers can even exhibit approximately a 32% decrease in attention to image regions relevant to queries (§5.1). This finding motivates us to guide the inference process of LVLMs for non-English queries by leveraging the cross-modal attention patterns in English scenarios, as LVLMs are typically well-trained on large English image-text data and perform best in English.

To this end, we propose **Cross-Lingual Attention Intervention for Mitigating Multilingual Object Hallucination (CLAIM)** in LVLMs, a near training-free, plug-and-play method that is applicable during the inference stage. We first identify language-specific cross-modal attention heads, *i.e.*, the attention heads behaving quite differently for visual tokens in the same meaning queries across various languages. Next, we estimate language shift vectors for caption queries of images from English to the target language. During the inference stage, we apply shift vectors to intervene in attention outputs of these heads to align with English visual perception capabilities, reducing the likelihood of multilingual object hallucination.

Experiments conducted on LLaVA-1.5 (Liu et al., 2024a) and Qwen-VL-Chat (Bai et al., 2023b) demonstrate that CLAIM results in an average improvement of 13.56% on the POPE (Li et al., 2023) benchmark and 21.75% on the hallucination subsets of the MME (Fu et al., 2023).

Our contributions are summarized as follows:

- We reveal significant cross-modal attention divergence across languages in LVLMs.
- We propose CLAIM, a novel inference-time method that aligns non-English attention patterns with English, mitigating multilingual object hallucination with low cost.
- We analyze LVLMs' attention patterns in multilingual scenarios, highlighting the role of intermediate layers in cross-modal inference.

2 Related Work

Multilingual Large Vision-Language Models Leveraging the advanced capabilities of Large Language Models (LLMs) (Touvron et al., 2023a; Chiang et al., 2023; Qin et al., 2025), Large Vision-Language Models (LVLMs) (Yin et al., 2023; Wu et al., 2023) integrate visual encoders (Dosovitskiy, 2020; Radford et al., 2021) and feature projectors, allowing them to process and generate content from both visual and textual inputs. Built on the strong multilingual language model LLaMA-2 (Touvron et al., 2023b), which is trained on a diverse multilingual corpus, LLaVA-1.5 (Liu et al., 2024a) is inherently a multilingual LVLM. Qwen-VL-Chat (Bai et al., 2023b) is another multilingual LVLM with an English-centric design, trained on a large corpus of Chinese-language data and built upon Qwen (Bai et al., 2023a). Existing research (Geigle et al., 2023; Andersland, 2024; Maaz et al., 2024) employ training-based approaches to enhance the multilingual capabilities of LVLMs. Despite significant progress, LVLMs still struggle with multilingual object hallucination, limiting their global applicability in diverse countries and languages.

Hallucination in LVLMs LVLMs often generate text outputs that are inconsistent with the visual input, a issue commonly referred to as the hallucination phenomenon. To mitigate hallucination, some methods focusing on the training phase utilize instruction (Liu et al., 2023), reinforcement learning with human/AI feedback (Yu et al., 2024), or model structure enhancement (Chen et al., 2024a). Another line of methods (Leng et al., 2024; Chen et al., 2024c; Zhong et al., 2024; Huang et al., 2024) reduce the likelihood of hallucination by performing conservative decoding on the original inputs and the inputs with disturbed contents. However, the above approaches proposed for mitigating hallucination only focus on their effectiveness in English. MHR (Qu et al., 2024) first attempts to mitigate multilingual object hallucination by SFT and DPO. In this paper, we propose a novel method for mitigating multilingual object hallucination without datasets construction and training.

3 Methodology

In this section, we first introduce the overall process of CLAIM in Figure 2, followed by the preliminary for the attention mechanism of LVLMs. Then, we describe the three parts of CLAIM in detail.

3.1 Preliminary

Modern LVLMs (Bai et al., 2023b; Liu et al., 2024a) typically comprise three key components: a visual encoder, a feature projector, and a language decoder. Specifically, the visual encoder first transforms the input visual image into visual features. The feature projector then maps these features to the input space of the language decoder, producing the visual embeddings $\boldsymbol{P} = \{p_i\}_{i=0}^{n}$, where p_i represents the visual embedding corresponding to the

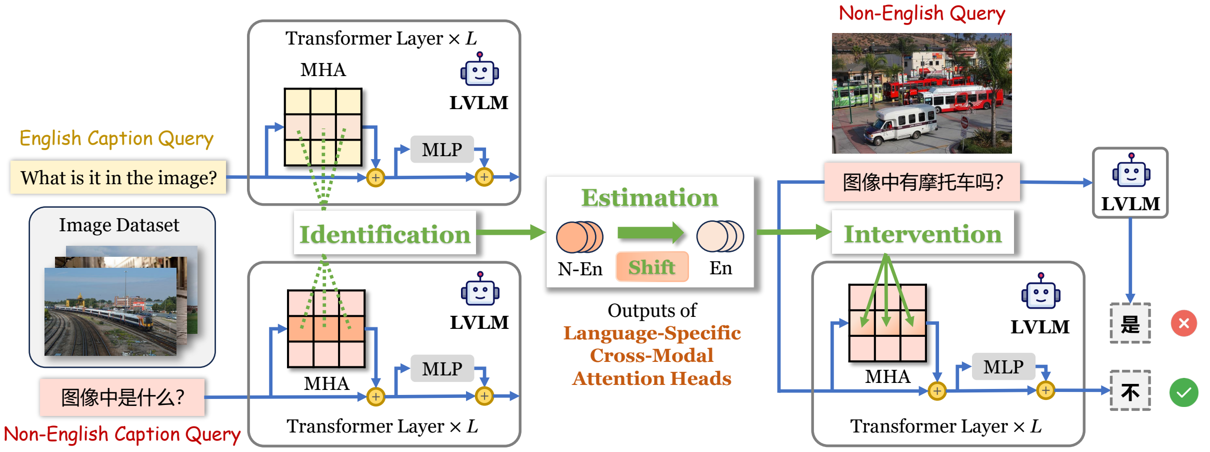


Figure 2: Overview of our proposed CLAIM method. A block of MHA in the figure represents a attention head. CLAIM intervene in identified language-specific cross-modal attention heads using estimated language shift vectors. **(1) Identification of Language-Specific Cross-Modal Attention Heads** §3.2: We train probes to identify the language-specific cross-modal attention heads, which exhibit significantly different behavior across languages associated with visual perception. **(2) Estimation of Language Shift Vectors** §3.3: We estimate the language shift vectors in attention outputs from English to the target non-English language for identical images queried with captions. **(3) Intervention during Inference** §3.4: During inference, we apply language shift vectors to intervene in the language-specific cross-modal attention heads for mitigating multilingual object hallucination.

i-th image patch, and n denotes the total number of patches. Similarly, the textual input is mapped to textual embeddings $\boldsymbol{T} = \{t_i\}_{i=0}^{m}$, where t_i corresponds to the i-th textual token, and m represents the number of text tokens. The visual and textual embeddings are then concatenated to form the input embeddings $\boldsymbol{X} = [\boldsymbol{P}, \boldsymbol{T}]$ for the language decoder.

During the forward pass of the language decoder, the input embeddings $\boldsymbol{X}$ serve as the hidden states for the first self-attention layer (Vaswani, 2017). The h-th attention head in the l-th self-attention layer applies linear transformations to project the hidden states into queries $\boldsymbol{Q}_h^l \in \mathbb{R}^{e \times d}$, keys $\boldsymbol{K}_h^l \in \mathbb{R}^{e \times d}$, and values $\boldsymbol{V}_h^l \in \mathbb{R}^{e \times d}$. Here, $e = n+m$ and d denotes the head-specific hidden dimension. The attention scores $\boldsymbol{A}_h^l \in \mathbb{R}^{e \times e}$ are then calculated based on $\boldsymbol{Q}_h^l$ and $\boldsymbol{K}_h^l$ as follows:

$$\widetilde{\boldsymbol{A}}_h^l = \text{softmax}(\boldsymbol{A}_h^l + \boldsymbol{M}), \boldsymbol{A}_h^l = \frac{\boldsymbol{Q}_h^l {\boldsymbol{K}_h^l}^T}{\sqrt{d}}, \quad (1)$$

$$\boldsymbol{M}[i,j] = \begin{cases} 0 & \text{if } j \leq i \\ -\infty & \text{if } j > i \end{cases} \quad (2)$$

where $\boldsymbol{M}$ is the causal mask matrix. The attention weights $\widetilde{\boldsymbol{A}}_h^l$ estimate the relevance of each token, which are used to reweight the values $\boldsymbol{V}_h^l$ from each token, producing the attention outputs $\boldsymbol{O}_h^l \in \mathbb{R}^{e \times d}$,

$$\boldsymbol{O}_h^l = \widetilde{\boldsymbol{A}}_h^l \boldsymbol{V}_h^l. \quad (3)$$

At each layer, the hidden states pass through multi-head attention (MHA), which comprises H independent attention heads, each performing separate linear transformations. Specifically, the MHA mechanism can be formulated as:

$$\boldsymbol{X}^{l+1} = \boldsymbol{X}^l + \sum_{h=1}^{H} \boldsymbol{O}_h^l \boldsymbol{W}_h^l, \quad (4)$$

where $\boldsymbol{W}_h^l \in \mathbb{R}^{d \times Hd}$ maps d-dimensional attention outputs of heads into hidden state representations, which are then fed into a standard multilayer perceptron (MLP) for further processing. Finally, the hidden state of the last token is decoded into a next-token prediction distribution.

3.2 Identification of Language-Specific Cross-Modal Attention Heads

Since LVLMs generate tokens in an auto-regressive manner, our method focuses on the attention matrices of the last input token, $\boldsymbol{A}_h^l[e]$, which aggregates the most comprehensive visual and textual information. We mask $\boldsymbol{A}_h^l[e]$ to exclude attention toward all textual tokens, allowing us to identify the language-specific cross-modal attention heads, which capture the variations in cross-modal attention patterns where semantically equivalent text in different languages attends to the same image. If we do not exclude the attention from the last input token to preceding text tokens, it would introduce substantial text-language-specific features, leading to the identification of attention heads specialized for the input text's language, which diverges from the primary motivation and purpose of our work.

$$\widehat{\boldsymbol{M}}[i,j] = \begin{cases} 0 & \text{if } j \leq i \\ -\infty & \text{if } j > i \text{ or } (i = e \text{ and } j > n) \end{cases} \tag{5}$$

$$\widehat{\boldsymbol{O}}_h^l = \text{softmax}(\boldsymbol{A}_h^l + \widehat{\boldsymbol{M}})\boldsymbol{V}_h^l. \tag{6}$$

For each image $\boldsymbol{P}_i$, we construct caption queries in both English and the target non-English language. A caption query refers to a request where the goal is to generate a textual description (a caption) for a given image which designed to stimulate LVLMs' visual perception capabilities and contributes to mitigating multilingual object hallucination. The English query, T_{en}, is "What is it in the image?", while T_{tgt} is its translation into the target language. Then, both queries are fed into the model along with the image for the standard inference process, deriving $\boldsymbol{x}_i \in \left\{\widehat{\boldsymbol{O}}_{i,h}^{en,l}[e_i], \widehat{\boldsymbol{O}}_{i,h}^{tgt,l}[e_i]\right\}$.

Probe (Li et al., 2024) f_h^{l*} is a binary classifier, trained to predict language labels based on $\boldsymbol{x}_i$. The language labels $y_i \in \{1, -1\}$ corresponds to English and the target language respectively. We train probes using B samples for each attention head H_h^l. Finally, we evaluate probes using test samples, identifying language-specific cross-modal attention heads for Top-K classification accuracy. The formulas are summarized as:

$$f_h^{l*} = \arg\min_{f_h^l} \sum_{i=1}^{B} \mathcal{L}\left(f_h^l(\boldsymbol{x}_i), y_i\right), \tag{7}$$

$$H_s = \{H_h^l \mid H_h^l \in TopK(\text{Acc}(f_h^{l*}))\}, \tag{8}$$

where $\mathcal{L}$ is the loss function of probes and H_s is the set of language-specific cross-modal attention heads, K denoted as the number of selected heads.

3.3 Estimation of Language Shift Vectors

Given the sets $\{(T_{en}, P_i)\}_{i=1}^B$ and $\{(T_{tgt}, P_i)\}_{i=1}^B$, we derive $\{\boldsymbol{O}_{i,h}^{en,l}\}_{i=1}^B$ and $\{\boldsymbol{O}_{i,h}^{tgt,l}\}_{i=1}^B$ respectively through the standard inference process of the model, estimating the language shift vectors $\boldsymbol{S}_h^l$:

$$\boldsymbol{S}_h^l = \frac{1}{B}\sum_{i=1}^{B}\left(\boldsymbol{O}_{i,h}^{en,l}[e_i] - \boldsymbol{O}_{i,h}^{tgt,l}[e_i]\right). \tag{9}$$

The shifts estimate the attention disparities between English and the target language for visual perception alignment. Notably, we do not use $\widehat{\boldsymbol{O}}_{i,h}^{en,l}, \widehat{\boldsymbol{O}}_{i,h}^{tgt,l}$ to estimate language shift vectors as these matrices are not directly derived from the standard inference process. Instead, in order to preserve the original representation space of the model, we opt to utilize $\boldsymbol{O}_{i,h}^{en,l}$ and $\boldsymbol{O}_{i,h}^{tgt,l}$ providing more reliable shifts. As the visual and textual representations in $\boldsymbol{O}_h^l$ are not orthogonal, the language shift vectors inherently capture multimodal information understanding which aids in mitigating multilingual object hallucination.

3.4 Intervention during Inference

Finally, we apply language shift vectors to intervene in language-specific cross-modal attention heads during inference in non-English queries:

$$\boldsymbol{X}^{l+1} = \boldsymbol{X}^l + \sum_{h=1}^{H}(\boldsymbol{O}_h^l + \mathbb{I}_h^l \alpha \boldsymbol{S}_h^l)\boldsymbol{W}_h^l. \tag{10}$$

$\mathbb{I}_h^l$ is an indicator function, which is 1 if $H_h^l \in H_s$ and 0 otherwise, and α denotes the intensity of the intervention. After intervention, LVLMs leverage their strongest English visual perception proficiency even when processing non-English queries. Since the intervention are pre-computed, CLAIM hardly incurs additional latency during the inference stage. We estimate the inference speed and discuss the results in Appendix D.

4 Experiment

4.1 Datasets

POPE POPE (Li et al., 2023) is designed to evaluate object hallucination in the VQA paradigm. It queries LVLMs about the presence of specific objects in a given image while maintaining a balanced 1:1 ratio between existent and non-existent objects. The benchmark employs three distinct sampling strategies for negative samples - random, popular, adversarial - with their difficulty levels increasing in that order. POPE integrates data from three major repositories: MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019). Evaluation is conducted using accuracy as the primary metric. As the original POPE benchmark is available only in English, we translating all queries into multiple languages using Google Translate and meticulously refine the translation results to maintain superior benchmark quality. Since the text is fairly simple, we directly use Google Translate and find it feasible to verify the translation quality, as detailed in Appendix B.

Dataset	Setup	Method	LLaVA-1.5							Qwen-VL-Chat						
			En	Zh	Es	Ru	Pt	Bg	Avg.	En	Zh	Es	Ru	Hi	De	Avg.
COCO	Random	*Baseline*	88.50	81.00	63.03	72.33	78.97	72.23	73.51	86.63	84.57	68.13	76.83	56.97	77.53	72.81
		VCD	-	81.47	67.40	73.33	78.07	72.47	74.55	-	84.70	72.17	75.37	48.37	78.27	71.78
		Ours	-	86.50	87.33	87.50	85.40	80.50	85.45	-	88.03	86.93	82.60	67.57	88.07	82.64
	Popular	*Baseline*	87.43	83.07	62.93	69.20	82.03	71.17	73.68	85.67	83.40	68.03	75.73	62.20	77.20	73.31
		VCD	-	83.20	67.20	70.17	80.10	69.47	74.03	-	82.70	72.70	74.20	57.40	76.70	72.74
		Ours	-	88.00	87.53	84.03	86.13	80.07	85.15	-	85.47	85.90	80.63	69.23	86.43	81.53
	Adversarial	*Baseline*	85.20	73.40	62.87	66.07	73.70	65.93	68.39	83.90	80.87	67.53	72.80	63.10	75.90	72.04
		VCD	-	74.27	67.30	66.47	73.63	66.33	69.60	-	79.40	71.97	70.83	54.73	75.03	70.39
		Ours	-	77.67	83.27	79.43	79.67	74.57	78.92	-	82.33	80.70	77.53	69.70	81.50	78.35
OKVQA	Random	*Baseline*	91.00	76.13	63.40	70.30	75.03	69.23	70.82	88.47	85.10	67.70	74.63	59.53	75.17	72.43
		VCD	-	77.93	68.17	72.17	75.67	70.33	72.85	-	84.30	72.23	73.67	49.07	75.33	70.92
		Ours	-	86.80	85.97	85.63	83.87	80.63	84.58	-	88.13	86.03	83.47	68.83	86.13	82.52
	Popular	*Baseline*	86.97	74.50	63.27	69.40	77.03	67.73	70.39	88.70	85.60	67.90	75.43	58.87	75.27	72.61
		VCD	-	77.57	67.10	68.40	77.03	68.20	71.66	-	84.17	71.57	73.30	52.40	75.63	71.41
		Ours	-	85.00	83.40	81.87	84.27	77.97	82.50	-	88.10	85.70	83.90	68.50	86.03	82.45
	Adversarial	*Baseline*	79.57	64.40	62.37	63.93	67.97	63.03	64.34	82.40	79.97	66.80	72.50	60.17	72.30	70.35
		VCD	-	68.97	65.17	63.90	68.67	64.07	66.16	-	78.77	69.00	69.63	50.80	71.90	68.02
		Ours	-	77.70	75.40	76.17	75.67	71.97	75.38	-	81.10	77.20	78.73	67.73	78.87	76.73
GQA	Random	*Baseline*	89.47	77.03	63.83	70.23	74.87	68.07	70.81	87.23	82.53	73.03	74.00	65.10	74.30	73.95
		VCD	-	78.00	68.33	71.47	73.63	70.20	72.33	-	83.63	76.10	73.43	55.50	80.20	73.77
		Ours	-	84.27	85.13	83.57	82.53	82.67	83.63	-	85.17	79.93	81.90	62.03	80.40	77.89
	Popular	*Baseline*	83.90	69.60	64.03	64.53	71.53	63.40	66.62	85.80	81.90	72.17	75.07	59.20	70.40	71.93
		VCD	-	73.60	67.50	65.23	71.87	66.20	68.88	-	82.20	73.73	74.17	54.70	76.87	72.33
		Ours	-	81.57	78.50	77.93	79.60	80.57	79.63	-	84.43	80.60	82.60	61.30	78.07	77.40
	Adversarial	*Baseline*	81.17	63.17	63.30	64.10	67.90	62.10	64.11	82.63	78.97	69.73	73.07	61.73	71.63	71.52
		VCD	-	67.33	66.97	63.87	67.83	63.37	65.87	-	79.57	73.47	71.83	54.97	76.17	71.20
		Ours	-	77.57	75.93	76.27	75.23	76.47	76.29	-	79.97	76.77	79.53	62.47	77.63	75.27
MME	Existence	*Baseline*	190.0	175.0	130.0	145.0	155.0	125.0	146.0	185.0	190.0	135.0	140.0	106.7	195.0	153.3
		VCD	-	180.0	130.0	145.0	150.0	128.3	146.7	-	185.0	155.0	155.0	78.30	195.0	153.7
		Ours	-	195.0	175.0	185.0	175.0	180.0	182.0	-	190.0	155.0	185.0	128.3	195.0	170.7
	Count	*Baseline*	155.0	70.00	55.00	58.30	80.00	85.00	69.67	150.0	130.0	143.3	113.3	60.00	136.7	116.7
		VCD	-	73.30	73.30	53.30	61.70	105.0	73.33	-	140.0	131.7	100.0	80.00	128.3	116.0
		Ours	-	125.0	130.0	110.0	130.0	135.0	126.0	-	148.3	153.3	120.0	111.7	137.0	134.0
	Color	*Baseline*	165.0	80.00	135.0	75.00	110.0	80.00	96.00	180.0	170.0	150.0	153.3	103.3	165.0	148.3
		VCD	-	95.00	145.0	85.00	130.0	88.30	108.7	-	150.0	165.0	146.7	93.30	160.0	143.7
		Ours	-	120.0	155.0	125.0	150.0	108.3	131.7	-	170.0	165.0	158.3	90.00	165.0	149.7
	Position	*Baseline*	118.3	53.30	63.30	55.00	50.00	46.70	53.67	131.7	63.30	120.0	93.30	45.00	105.0	85.33
		VCD	-	48.30	93.00	60.00	51.70	56.70	61.93	-	78.30	101.7	101.7	43.30	93.30	83.67
		Ours	-	66.70	78.30	85.00	86.70	58.30	75.00	-	63.30	116.7	103.3	55.00	106.0	88.86
	Total Scores	*Baseline*	628.3	378.3	383.3	333.3	395.0	336.7	365.3	646.7	553.3	548.3	500.0	315.0	601.7	503.7
		VCD	-	396.7	441.3	343.3	393.3	378.3	390.6	-	553.3	556.7	503.3	295.0	576.7	497.0
		Ours	-	506.7	538.3	505.0	541.7	481.7	514.7	-	571.7	590.0	566.6	385.0	603.0	543.3

Table 1: Main results on POPE from COCO, OKVQA, GQA and the hallucination subsets of MME.

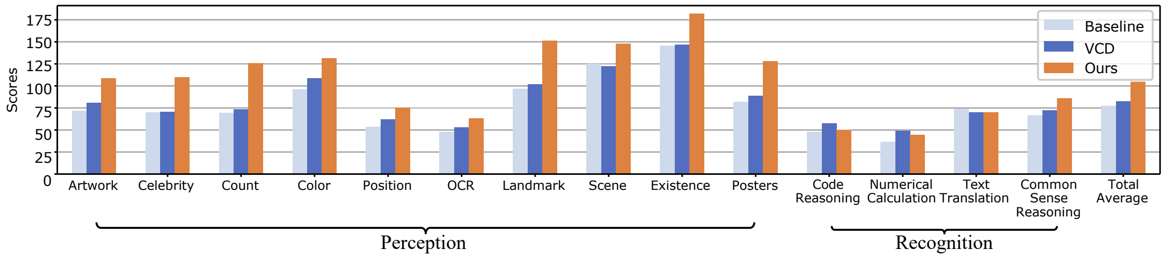


Figure 3: Average scores for LLaVA-1.5 across five languages on the MME full dataset.

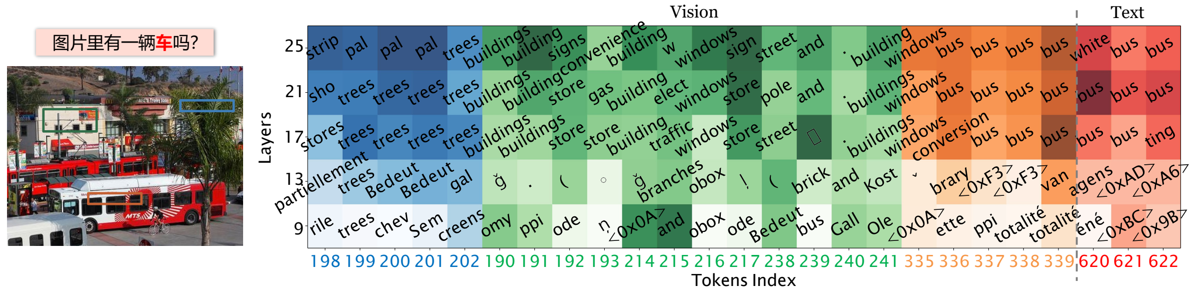


Figure 4: Logit lens observation for interpreting LLaVA-1.5 in multilingual scenarios. The depth of block color for the i-th token at layer l indicates the magnitude of its contribution to the logits of the final predicted token. The color represents the corresponding tokens in image or text. The query means "Is there a car in the image?".

MME The MME dataset (Fu et al., 2023) serves as a comprehensive benchmark for evaluating LVLMs, including 14 subtasks designed to assess both the perceptual and cognitive abilities of LVLMs. Performance is measured based on the sum of accuracy scores across individual questions and images. Following Leng et al. (2024), in addition to adapting the full dataset, we focus specifically on the existence and count subsets for object-level hallucination evaluation, as well as the position and color subsets for attribute-level assessment. Similar to POPE, we translate MME into seven languages using Google Translate and correct the mistakes by human check.

4.2 Models and Implementation Details

Following prior research, we adopt the widely used LLaVA-1.5-7b (Liu et al., 2024a) and Qwen-VL-Chat (Bai et al., 2023b) as our baseline LVLMs. More LVLM results can be found in Appendix C. Since no existing training-free methods tailored to mitigating multilingual object hallucination, we employ the widely used method, VCD (Leng et al., 2024), as a strong baseline for comparison. We sample 1,000 images from the COCO-2017 training dataset to complete identification and estimation, discussing the impact of training set size in Appendix F. To determine optimal hyperparameter values for α (intervention intensity) and K (the number of heads involved in the intervention), we employ a sequential optimization approach. Additional details are provided in Appendix A.

4.3 Main Results

Results on POPE. **(1) CLAIM effectively mitigates object-level hallucination** by aligning the strong visual perception capability used for processing English queries with that used for processing non-English queries. As shown in Table 1, the intervention achieves an average improvement of 17.5% over the baseline on LLaVA-1.5 and 9.8% on Qwen-VL-Chat across various languages and settings. **(2) CLAIM enhances performance across both low- and high-resource non-English languages**. This improvement can be attributed to its robust ability to facilitate cross-lingual visual perception capability alignment almost regardless of the language's resource availability. **(3) The intervention yields improvements across datasets with different distributions**, suggesting that the intervention represents a generalizable direction to mitigating multilingual object hallucination rather than merely tailored to a specific dataset.

Results on MME. This subset extends beyond POPE's scope, encompassing both object-level and attribute-level hallucinations **(1) CLAIM effectively reduces both object-level and attribute-level hallucination.** As shown in Table 1, CLAIM achieves an average improvement of 40.9% on LLaVA-1.5 and 7.9% on Qwen-VL-Chat over the baseline across various languages, outperforming VCD. Specifically, CLAIM not only mitigates object-level hallucination, as evidenced by the results on the existence and count subsets, but also mitigates attribute-level hallucination, as demonstrated by the color and position subsets. Detailed results can be found in Appendix E. **(2) CLAIM could generally facilitate cross-lingual visual perception capability alignment** by intervening the attention patterns, enabing LVLMs to transfer their English proficiency across various tasks in multilingual queries. Illustrated in Figure 3, the intervention significantly enhances perception-based tasks and generalizes well to cognitive reasoning tasks, as strong image perception serves as the foundation for cognitive processing. Meanwhile, our method primarily activates attention heads associated with perception, which could unintentionally affect the reasoning pathways of LVLMs.

5 Analysis and Discussion

5.1 Multilingual Attention Differences

As shown in Figure 1, when queried in different languages about an object's presence, the model exhibits distinct attention patterns across image regions. In order to quantitatively analyze multilingual cross-modal attention differences, we conduct a statistical experiment to validate our motivation. Specifically, using bounding box annotations from the POPE-COCO dataset, we localize object regions and compute the sum of attention weights within ground-truth bounding boxes for queries in different languages, obtaining $A_b(l, \mathcal{B})$.

$$A_b(l, \mathcal{B}) = \frac{n}{H|\mathcal{B}|} \cdot \sum_{h=1}^{H} \left(\frac{\sum_{j \in \mathcal{B}} \widetilde{A}_h^l(e, j)}{\sum_{j=1}^{n} \widetilde{A}_h^l(e, j)} \right) \quad (11)$$

$\mathcal{B}$ denotes the set of patches included in bounding boxes. The first fraction serves as a normalization factor, eliminating the influence of varying bounding box sizes associated with different queries and averaging H attention heads in each layer.

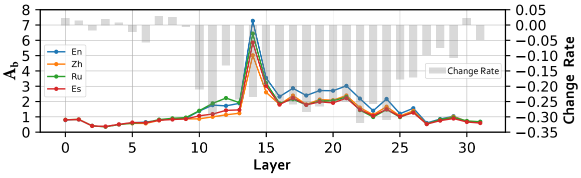


Figure 5: A_b per layer of LLaVA-1.5 across four languages, and the per-layer average change rate of non-English languages A_b relative to English.

As shown in Figure 5, in the initial layers of the model (before layer 10), the model pays little attention to image regions related to queries, and there is almost no difference between multilingual queries, indicating that the model is performing basic global understanding of the image at this stage. In the middle layers (10-25), the attention to key regions in the image for non-English languages decreases significantly compared to English, suggesting that the model exhibits disparities in perceptual capabilities under different language queries. Notably, a peak appears at layer 14, even though the model's total attention weights for visual tokens are low at this stage (shown in Figure 7 (a)). This may indicate that the model has noticed image regions relevant to queries during its internal reasoning process, performing fine-grained perception.

5.2 Multilingual Inference in LVLMs

In order to better illustrate the mechanism behind CLAIM and improve the interpretability of LVLMs in multilingual scenarios, by leveraging the logit lens (Nostalgebraist, 2020) method, which aims to decode the hidden states of the language decoder at various layers, we investigate the internal inference mechanism of LVLMs and uncover how they process and integrate multimodal information, particularly in non-English queries. The internal inference process of LLaVA-1.5 is illustrated in Figure 4 as a case study.

For multilingual VQA tasks, English-centric LVLMs face the dual challenge of not only bridging the modality gap between visual and textual information but also mapping non-English queries into the English semantic space to ensure accurate responses. **In middle layers, visual tokens are often decoded into their corresponding English concepts**. Similarly, when processing Chinese queries, the model maps them to the English semantic space at these intermediate layers such as the layer 17, consistent with the findings (Wendler et al., 2024) in multilingual LLMs. During pretraining, the alignment of vision-text modality relies heavily on English corpora, which guides LVLMs toward interpreting images through an English-centric pathway. **LVLMs interpret the important entities in the query based on the critical information from the image at intermediate stage**. The original semantic meaning of "car" in the query (such as the 620th token) is enriched to "bus" under the influence of the image information at the layer 21.

5.3 Analysis of Intervention

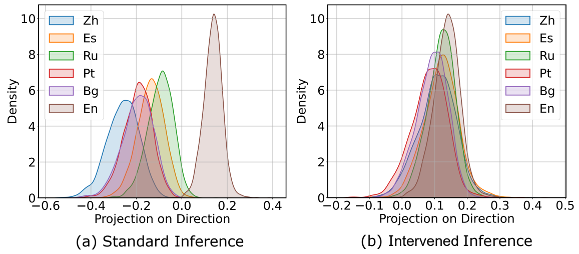
In order to elucidate the underlying reasons for the efficacy of CLAIM and to substantiate the robustness of our methodology, we conduct a further investigation with two questions.

Does the intervention truly facilitate the cross-lingual alignment of cross-modal attention distribution? Illustrated in Figure 6, under standard inference, the projection values of English and non-English vectors are distinctly separated at the zero point, indicating that the cross-modal attention patterns of LVLMs exhibit significant misalignment for identical images depending on the language used in the same meaning query. After intervention, the distributions of non-English languages shift closer to that of English, with more pronounced density peaks. This alignment supports the hypoth-

Setup	Method	LLaVA-1.5						Qwen-VL-Chat					
		Zh	Es	Ru	Pt	Bg	Avg.	Zh	Es	Ru	Hi	De	Avg.
Random	*Baseline*	81.00	63.03	72.33	78.97	72.23	73.51	84.57	68.13	76.83	56.97	77.53	72.81
	Mono-Shift	86.50	86.70	87.40	84.50	76.77	84.37	88.03	74.87	80.67	67.27	83.03	78.77
	Multi-Shift	85.80	87.00	87.87	83.60	79.80	84.81	86.70	86.40	81.37	63.30	87.97	81.15
	Specific-Shift	86.50	87.33	87.50	85.40	80.50	85.45	88.03	86.93	82.60	67.57	88.07	82.64
Popular	*Baseline*	83.07	62.93	69.20	82.03	71.17	73.68	83.40	68.03	75.73	62.20	77.20	73.31
	Mono-Shift	88.00	86.90	83.00	85.30	79.10	84.46	85.47	74.77	78.80	68.07	81.73	77.77
	Multi-Shift	87.67	87.23	83.33	86.17	79.13	84.71	84.30	85.03	80.37	68.93	85.43	80.81
	Specific-Shift	88.00	87.53	84.03	86.13	80.07	85.15	85.47	85.90	80.63	69.23	86.43	81.53
Adversarial	*Baseline*	73.40	62.87	66.07	73.70	65.93	68.39	80.87	67.53	72.80	63.10	75.90	72.01
	Mono-Shift	77.67	82.63	78.13	78.33	72.60	77.87	82.33	73.83	76.00	69.30	79.57	76.21
	Multi-Shift	76.70	82.13	79.53	78.10	74.83	78.66	81.70	80.50	76.57	69.97	81.17	77.98
	Specific-Shift	77.67	83.27	79.43	79.67	74.57	78.92	82.33	80.70	77.53	69.70	81.50	78.35

Table 2: Evaluation results of different intervention estimation approaches on POPE-COCO. Green means the best perfomance while gray means the second-best results.

esis that CLAIM effectively mitigates multilingual object hallucination by reinforcing cross-lingual consistency in visual perception attention.



(a) Standard Inference (b) Intervened Inference

Figure 6: Kernel density estimate plot of cross-modal attention outputs across languages of LLaVA-1.5 before and after multilingual intervention. The x-axis represents the inner product between attention outputs and the normal vector of the hyperplane, while the y-axis indicates the density of samples occurring at x value.

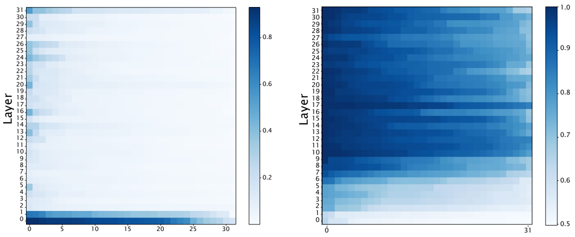
Does a unified multilingual intervention (Multi-Shift) work? In our main experiment, we pair English with each non-English language individually, estimating the *Specific-Shift* for each pair and selecting specialized attention heads to precisely align each non-English languages with English proficiency. To examine the impact of multilingual interactions, we construct a mixed set that includes all non-English languages paired with English, estimating the *Multi-Shift* between English and the entire non-English group. As shown in Table 2, the *Multi-Shift* demonstrates moderate improvements, even outperforming CLAIM in some subsets by integrating shared linguistic features. Additionally, to further validate the generalizability of the intervention, we apply the *Mono-Shift* intervention, derived from the English-Chinese pair, to other languages and observe performance improvements. It indicates that the cross-lingual intervention can generalize well to unseen non-English languages.

5.4 Analysis of Attention Heads

LVLMs extract and interpret cross-modal information within language-specific cross-modal attention heads. As shown in Figure 7 (b), theses attention heads are predominantly located in the intermediate layers of the model, particularly at layers 10-17, suggesting that cross-modal integration occurs primarily at this stage. Beyond this, we examine how different layers influence the final prediction. In Figure 4, we observe that around layer 21, visual and textual tokens—such as the 339th and 620th tokens, which correspond to the semantics of "bus"—have a significant impact on the logits of the final prediction. Notably, attention heads in these layers exhibit high classification accuracy, indicating their crucial role in linguistic visual perception and understanding. Conversely, Figure 7 (a) shows that in the early layers (0 and 1), the attention weights of the last input token to visual tokens are strongest. However, classification accuracy remains very low in these heads, suggesting that while LVLMs engage in basic image feature extraction at the initial layers, they contribute minimally to linguistic understanding.

5.5 Impact of Hyperparameters

CLAIM is primarily governed by two key hyperparameters: the intervention intensity α and the number of heads K involved in the intervention. Illustrated in Figure 8, the ablation experiments vary one parameter while keeping the other fixed,



(a) Attention Weights to Visual Tokens

(b) Classification Accuracy of Probes

Figure 7: Heatmap of (a) sum attention weights of the last input token toward all visual tokens and (b) the classification accuracy of probes of LLaVA-1.5 across all 32x32 multi-heads, sorted row-wise by value.

yielding several key insights. When α is too small, the intervention is insufficient, resulting in suboptimal improvements. However, an excessively large α imposes an overly strong intervention, disrupting the LVLMs' capabilities. For the hyperparameter K, we observe that a small K leads to inadequate intervention in language-specific cross-modal attention heads, reducing effectiveness. On the other hand, a large K introduces unnecessary interference by affecting attention heads that encode irrelevant information, ultimately degrading performance. Overall, CLAIM achieves performance improvements across a wide range of hyperparameter settings, demonstrating strong robustness to hyperparameter selection.

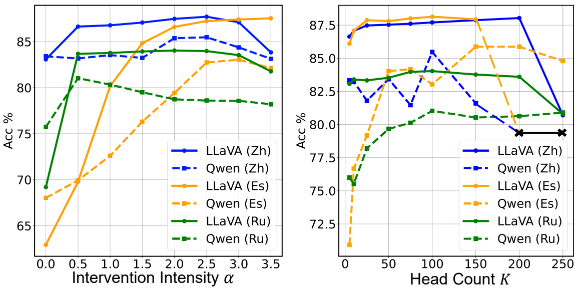


Figure 8: Impact of hyperparameters α and K on the Accuracy for LLaVA-1.5 and Qwen-VL-Chat on the POPE-COCO popular subset. The "x" symbol indicates a value extraordinarily lower than normal.

6 Conclusion

In this paper, we propose **Cross-Lingual Attention Intervention for Mitigating Multilingual Object Hallucination (CLAIM)** in LVLMs, a near training-free method that aligns attention patterns across languages. Extensive evaluations on POPE and MME benchmarks demonstrate that CLAIM effectively mitigates multilingual object hallucination and generalizes well across languages and datasets. Further analysis reveals that attention discrepancies primarily occur in intermediate layers and LVLMs extract and interpret cross-modal information within language-specific cross-modal attention heads., providing deeper insights into multilingual LVLMs inference pathways.

7 Limitations

CLAIM requires distinguishing text and vision information within the attention mechanism to identify language-specific cross-modal attention heads, making it applicable only to LVLMs that treat visual and textual tokens equally in the language decoder. Additionally, our method requires access to the internal layers and representations of LVLMs, limiting its applicability to closed-source models. How to mitigate multilingual object hallucination as a plug-and-play tool for all LLMs, including those with restricted access, requires further investigation. Besides, since CLAIM aligns non-English attention patterns with English patterns, it may inadvertently reinforce English-centric biases rather than fostering truly multicultural comprehension.

8 Acknowledgments

Xiaocheng Feng is the corresponding author of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) (grant 62276078, U22B2059), the Key R&D Program of Heilongjiang via grant 2022ZX01A32, and the Fundamental Research Funds for the Central Universities (Grant No.HIT.OCEF.2023018).

References

Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *arXiv preprint arXiv:2403.06354*.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, and Chenliang Xu. 2024. Unveiling visual perception in language models: An attention head analysis approach. *arXiv preprint arXiv:2412.18108*.

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2024a. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Preprint*, arXiv:2405.15356.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. 2024b. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2411.15268*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024c. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

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

Corinna Cortes. 1995. Support-vector networks. *Machine Learning*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavas. 2023. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. *CoRR*, abs/2307.06930. ArXiv: 2307.06930.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. 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.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2024. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Shi Liu, Kecheng Zheng, and Wei Chen. 2025. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Muhammad Maaz, Hanoona Abdul Rasheed, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Tim Baldwin, Michael Felsberg, and Fahad Shahbaz Khan. 2024. PALO: A Polyglot Large Multimodal Model for 5B People. *CoRR*, abs/2402.14818. ArXiv: 2402.14818.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. LessWrong.

OpenAI. 2023. GPT-4. https://openai.com/gpt-4.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Xiaoye Qu, Mingyang Song, Wei Wei, Jianfeng Dong, and Yu Cheng. 2024. Mitigating multilingual hallucination in large vision-language models. *arXiv preprint arXiv:2408.00550*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, et al. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. *CoRR*, abs/2406.05967. ArXiv: 2406.05967.

Yin S., Fu C., Zhao S., Li K., Sun X., Xu T., and Chen E. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Florian Schneider and Sunayana Sitaram. 2024. M5 - A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4309–4345. Association for Computational Linguistics.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. PMLR.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Bai Z., Wang P., Xiao T., He T., Han Z., Zhang Z., and Shou MZ. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *Preprint*, arXiv:2311.16839.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. *arXiv preprint arXiv:2407.00569*.

Peipei Zhu, Xiao Wang, Lin Zhu, Zhenglong Sun, Wei-Shi Zheng, Yaowei Wang, and Changwen Chen. 2023. Prompt-based learning for unpaired image captioning. *IEEE Transactions on Multimedia*, 26:379–393.

A Implementation Details

In our experiments, we use greedy search to ensure reproducibility. We employ the experimental settings as default in the VCD code repository. All datasets used in this paper are licensed under a Creative Commons Attribution 4.0 License. We conduct extensive experiments on languages exhibiting varying performance levels across the two models. Probe is implemented as a linear Support Vector Machine (SVM) (Cortes, 1995), using the default `LinearSVC` API from `Scikit-learn` (Pedregosa et al., 2011). When calculating metrics for discriminative tasks, we consider "yes" and "no" as right and wrong labels. The output of LVLMs are considered correct if its meaning matches the label. We utilize ChatGPT (OpenAI, 2023) to assist us with coding and polishing the paper.

The hyperparameters under consideration included α and K. The hyperparameter tuning strategy employed in this study follows a sequential optimization approach. We conduct hyperparameter tuning exclusively on the popular subset of POPE-COCO, with the search space for K defined as $\{50, 100, 150, 200, 250, 300\}$ and for α as $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5\}$. Initially, K was fixed at a value of 100, while α was systematically adjusted to identify its optimal setting. Through this process, the optimal value for α was determined. Subsequently, α was fixed at this optimal value, and K was iteratively tuned. This led to the identification of the optimal value for K. Consequently, the optimal hyperparameter combination was established as shown in Figure 8. This stepwise optimization strategy ensures a focused and efficient exploration of the hyperparameter space, leading to the identification of the most effective parameter configuration for the model.

B Evaluation of Translation Quality

We sample 100 translated queries from the POPE-COCO random subset for each language and back-translate them into English using Google Translate. The back-translated English queries are then input into LVLMs to test whether their predictions align with those generated from the original English queries. High prediction consistency indicates that the translated data maintains superior benchmark quality. Prediction consistency: Zh-100%, Es-100%, Ru-100%, Pt-100%, Bg-100%, Hi-100%, De-100%. These results demonstrate the reliability of our constructed multilingual dataset.

C Results of LLaVA-NeXT

To comprehensively demonstrate the effectiveness of CLAIM, we also conduct experiments on an advanced LVLM, LLaVA-NeXT-7b (Liu et al., 2024b), with stronger English capability than LLaVA-1.5-7b, as shown in Table 3, also demonstrating strong effectiveness.

D Inference Speed

We evaluate Tokens Per Second (TPS) of LLaVA-1.5 on the popular subset of POPE-COCO using different methods, with the experiments conducted on the H100 GPUs. The results show that VCD leads to a significant decrease in inference speed. We attribute this slowdown to the fact that contrastive-decoding-based methods typically require multiple inference runs or involve substantial additional computations during the inference process. In contrast, CLAIM introduces almost no extra computational overhead during inference, further showing the advantages of our method.

Method	TPS	Acc (%)
LLaVA-1.5-7b	55.46 ×1.0	73.68
+VCD	20.67 ×0.4	74.03
+Ours	54.79 ×1.0	85.15

E Detailed Results on MME

In Table 4, we present the performance of LVLM baselines on the 14 tasks of the MME benchmark. The Existence and Count subsets assess object-level hallucination, while the Color and Position subsets focus on attribute-level hallucination. These four subsets form the hallucination evaluation set of MME, which, together with the remaining six categories—Artwork, Celebrity, OCR, Landmark, Scene, and Poster—collectively evaluate the LVLMs' perception capabilities. Besides, Code Reasoning, Numerical Calculation, Text Translation and Commonsense Reasoning evaluate the LVLMs' recognition capabilities. The deployment of CLAIM nearly consistently improves their perceptual competencies, indicating the effectiveness of CLAIM for cross-lingual visual perception alignment. Furthermore, The results on recognition-related tasks indicate that the application of CLAIM, while mitigating hallucination issues and augmenting perceptual capabilities, remains effective on some reasoning tests.

Dataset	Setup	Method	En	Zh	Es	Ru	Pt	Bg	Avg.
COCO	Random	*Baseline*	88.50	88.33	80.30	88.37	86.77	81.70	85.09
		VCD	-	85.83	69.47	84.53	81.97	75.67	79.49
		Ours	-	88.73	88.47	88.13	86.73	81.67	86.75
	Popular	*Baseline*	87.37	87.53	80.07	85.60	85.13	79.30	83.53
		VCD	-	85.43	69.20	82.07	82.93	75.67	79.06
		Ours	-	88.43	88.13	85.83	85.40	80.33	85.62
	Adversarial	*Baseline*	86.30	80.80	79.20	80.60	80.37	76.10	79.41
		VCD	-	77.23	68.10	76.93	76.27	70.30	73.77
		Ours	-	80.63	83.83	80.20	80.30	75.83	80.16
OKVQA	Random	*Baseline*	91.00	83.03	67.50	85.00	78.97	81.20	79.14
		VCD	-	82.20	69.27	82.23	79.37	75.70	77.75
		Ours	-	85.47	87.50	86.50	83.33	80.90	84.74
	Popular	*Baseline*	89.00	81.93	67.40	80.73	80.80	79.90	78.15
		VCD	-	81.23	69.33	78.33	79.77	74.27	76.59
		Ours	-	85.33	85.27	82.13	85.07	80.13	83.59
	Adversarial	*Baseline*	81.97	70.80	66.47	74.17	71.27	73.33	71.21
		VCD	-	71.57	66.97	71.67	70.83	69.63	70.13
		Ours	-	75.40	76.93	76.73	75.17	73.70	75.59
GQA	Random	*Baseline*	89.93	82.53	67.73	84.47	80.07	79.87	78.93
		VCD	-	81.03	68.60	81.57	77.63	74.53	76.67
		Ours	-	84.23	87.13	85.47	83.30	79.97	84.02
	Popular	*Baseline*	85.97	77.17	67.70	73.93	74.90	79.73	74.69
		VCD	-	76.80	69.53	73.60	75.23	74.20	73.87
		Ours	-	80.50	83.77	76.50	81.83	81.03	80.73
	Adversarial	*Baseline*	82.60	70.33	66.93	73.90	71.50	74.03	71.34
		VCD	-	69.57	67.50	72.90	71.57	69.33	70.17
		Ours	-	74.07	78.10	76.93	76.53	74.87	76.10

Table 3: Main results of LLaVA-NeXT on POPE from COCO, OKVQA, GQA.

F Impact of Training Size

In the main experiment, we sample 1,000 images from the COCO-2017 training dataset to identify the language-specific cross-modal attention heads and estimate language shift vectors. Of these, 80% are used as the training set for the probe, while the remaining 20% serve as the test set for the probe. Specifically, CLAIM achieves strong performance even with a minimal training set. We validate this on the popular subset of POPE-COCO (Chinese), as shown in Figure 9. Notably, most improvements are achieved with as few as N = 50 training samples. As the training set size increases, identification becomes more robust; however, an excessively large training set may introduce additional noise when estimating the mean of the language shift vectors.

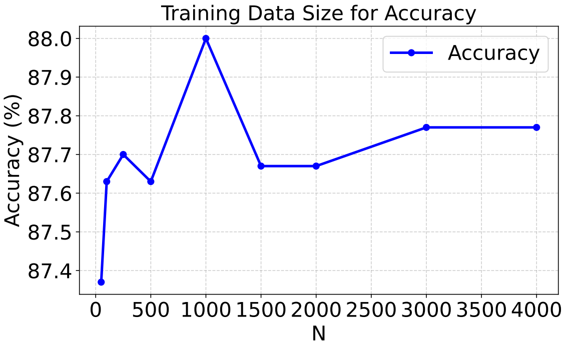


Figure 9: Impact of training data size on the Accuracy for LLaVA-1.5.

Task	Method	LLaVA-1.5							Qwen-VL-Chat						
		En	Zh	Es	Ru	Pt	Bg	Avg.	En	Zh	Es	Ru	Hi	De	Avg.
Existence	*Baseline*	190.0	175.0	130.0	145.0	155.0	125.0	146.0	185.0	190.0	135.0	140.0	106.7	195.0	153.3
	VCD	-	180.0	130.0	145.0	150.0	128.3	146.7	-	185.0	155.0	155.0	78.30	195.0	153.7
	Ours	-	195.0	175.0	185.0	175.0	180.0	182.0	-	190.0	155.0	185.0	128.3	195.0	170.7
Count	*Baseline*	155.0	70.00	55.00	58.30	80.00	85.00	69.67	150.0	130.0	143.3	113.3	60.00	136.7	116.7
	VCD	-	73.30	73.30	53.30	61.70	105.0	73.33	-	140.0	131.7	100.0	80.00	128.3	116.0
	Ours	-	125.0	130.0	110.0	130.0	135.0	126.0	-	148.3	153.3	120.0	111.7	137.0	134.0
Color	*Baseline*	165.0	80.00	135.0	75.00	110.0	80.00	96.00	180.0	170.0	150.0	153.3	103.3	165.0	148.3
	VCD	-	95.00	145.0	85.00	130.0	88.30	108.7	-	150.0	165.0	146.7	93.30	160.0	143.7
	Ours	-	120.0	155.0	125.0	150.0	108.3	131.7	-	170.0	165.0	158.3	90.00	165.0	149.7
Position	*Baseline*	118.3	53.30	63.30	55.00	50.00	46.70	53.67	131.7	63.30	120.0	93.30	45.00	105.0	85.33
	VCD	-	48.30	93.00	60.00	51.70	56.70	61.93	-	78.30	101.7	101.7	43.30	93.30	83.67
	Ours	-	66.70	78.30	85.00	86.70	58.30	75.00	-	63.30	116.7	103.3	55.00	106.0	88.86
Artwork	*Baseline*	121.8	86.25	57.75	73.00	81.25	61.25	71.90	135.3	152.8	105.8	70.00	65.75	130.5	105.0
	VCD	-	96.75	79.50	81.75	74.50	72.25	80.95	-	137.5	108.5	71.75	55.50	123.3	99.30
	Ours	-	117.5	118.0	103.5	107.3	98.25	108.9	-	146.0	110.0	115.0	83.00	133.8	117.6
Celebrity	*Baseline*	138.2	137.4	55.29	26.18	119.7	13.24	70.35	150.0	162.9	159.4	52.65	57.65	126.8	111.9
	VCD	-	139.7	79.12	23.82	99.41	10.59	70.53	-	160.9	158.5	67.35	52.35	121.2	112.1
	Ours	-	156.5	158.2	61.47	145.3	27.65	109.8	-	157.9	160.3	110.6	76.47	106.2	122.3
OCR	*Baseline*	125.0	65.00	50.00	55.00	15.00	55.00	48.00	102.5	80.00	115.0	62.50	32.50	132.5	84.50
	VCD	-	62.50	57.50	52.50	35.00	57.50	53.00	-	87.50	100.0	52.50	42.50	115.0	79.50
	Ours	-	80.00	82.50	57.50	32.50	65.00	63.50	-	87.50	100.0	77.50	50.00	135.0	90.00
Landmark	*Baseline*	165.3	113.3	53.75	81.50	127.3	108.5	96.90	172.8	179.0	123.3	78.25	61.50	163.3	121.1
	VCD	-	125.5	64.25	97.25	126.3	108.0	102.3	-	167.0	131.3	78.25	57.00	138.0	114.3
	Ours	-	147.8	155.5	146.5	159.3	148.3	151.5	-	178.3	129.8	160.8	88.25	167.5	144.9
Scene	*Baseline*	159.5	159.3	79.50	120.0	144.3	122.5	125.1	161.5	178.8	124.8	114.5	72.75	131.5	124.5
	VCD	-	149.8	87.30	119.5	137.5	118.0	122.4	-	162.3	122.8	108.0	71.50	128.8	118.7
	Ours	-	149.5	146.8	154.3	144.3	145.0	148.0	-	162.5	123.5	153.0	92.50	143.5	135.0
Poster	*Baseline*	143.5	106.1	67.35	57.14	110.2	68.71	81.90	173.1	156.8	142.9	116.7	76.19	134.4	125.4
	VCD	-	116.7	89.80	65.31	90.48	80.95	88.64	-	147.3	137.4	101.0	65.99	126.9	115.7
	Ours	-	133.7	156.8	97.62	129.9	121.8	128.0	-	161.6	150.3	138.8	109.9	137.8	139.7
Code Reasoning	*Baseline*	67.50	65.00	50.00	47.50	55.00	22.50	48.00	55.00	57.50	57.50	20.00	12.50	52.50	40.00
	VCD	-	67.50	57.50	72.50	62.50	27.50	57.50	-	42.50	57.50	37.50	35.00	45.00	43.50
	Ours	-	55.00	60.00	47.50	65.00	25.00	50.50	-	72.50	50.00	45.00	10.00	52.50	46.00
Numerical Calculation	*Baseline*	70.00	47.50	50.00	45.00	20.00	20.00	36.50	32.50	65.00	45.00	27.50	45.00	37.50	44.00
	VCD	-	75.00	50.00	62.50	37.50	20.00	49.00	-	80.00	60.00	37.50	45.00	45.00	53.50
	Ours	-	67.50	50.00	55.00	20.00	30.00	44.50	-	45.00	45.00	55.00	45.00	27.50	43.50
Text Translation	*Baseline*	70.00	77.50	50.00	70.00	120.0	52.50	74.00	155.0	60.00	115.0	110.0	17.50	65.00	73.50
	VCD	-	80.00	50.00	72.50	75.00	72.50	70.00	-	55.00	87.50	85.00	45.00	72.50	69.00
	Ours	-	95.00	57.50	50.00	90.00	57.50	70.00	-	87.50	122.5	110.0	50.00	72.50	88.50
Commonsense Reasoning	*Baseline*	124.3	72.14	64.29	61.43	72.86	64.29	67.00	125.0	92.14	100.7	50.71	33.57	74.29	70.28
	VCD	-	80.71	71.43	70.00	77.14	62.86	72.43	-	84.29	98.57	48.57	29.29	78.57	67.86
	Ours	-	90.00	87.14	86.43	87.14	68.57	85.86	-	91.43	107.1	80.00	39.29	84.29	80.43
Total Scores	*Baseline*	1813	1307	961.3	970.1	1261	925.2	1085	1909	1738	1638	1203	790.0	1650	1404
	VCD	-	1391	1128	1051	1209	1009	1157	-	1683	1619	1191	794.1	1571	1371
	Ours	-	1599	1611	1375	1522	1269	1475	-	1762	1689	1612	1029	1664	1551

Table 4: Detailed results on full subsets of MME.

G Comparison to PAI

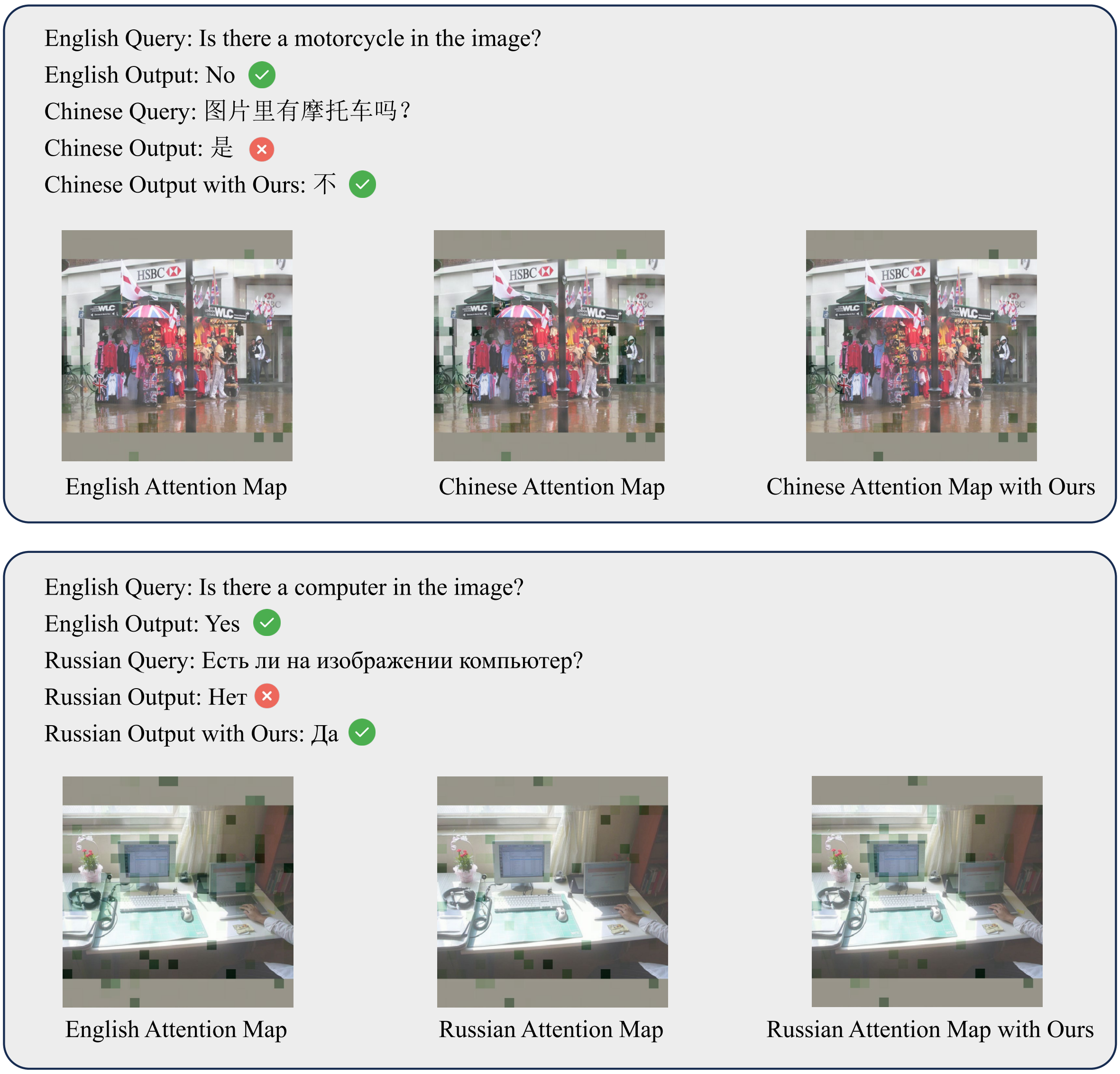
According to the suggestion of the anonymous reviewer, we conduct a comparison experiment with PAI (Liu et al., 2025), which intervenes on attention heads by leveraging their original direction. In Table 5, CLAIM still achieves best performance.

H Case Study

Illustrated in Figure 10, we present examples demonstrating how CLAIM effectively mitigates multilingual object hallucination. We compare the changes in the model's attention weight maps for the same image before and after applying CLAIM.

Setup	Method	Zh	Es	Ru	Pt	Bg	Avg.
Random	*Baseline*	81.00	63.03	72.33	78.97	72.23	73.51
	VCD	81.47	67.40	73.33	78.07	72.47	74.55
	PAI	80.10	67.97	83.50	79.13	71.10	76.36
	Ours	86.50	87.33	87.50	85.40	80.50	85.45
Popular	*Baseline*	83.07	62.93	69.20	82.03	71.17	73.68
	VCD	83.20	67.20	70.17	80.10	69.47	74.03
	PAI	83.90	67.73	79.17	82.90	71.90	77.12
	Ours	88.00	87.53	84.03	86.13	80.07	85.15
Adversarial	*Baseline*	73.40	62.87	66.07	73.70	65.93	68.39
	VCD	74.27	67.30	66.47	73.63	66.33	69.60
	PAI	76.60	67.43	75.20	75.53	67.73	72.50
	Ours	77.67	83.27	79.43	79.67	74.57	78.92

Table 5: Comparison to PAI on POPE of LLaVA-1.5.



English Query: Is there a motorcycle in the image?

English Output: No ✓

Chinese Query: 图片里有摩托车吗？

Chinese Output: 是 ✗

Chinese Output with Ours: 不 ✓

English Query: Is there a computer in the image?

English Output: Yes ✓

Russian Query: Есть ли на изображении компьютер?

Russian Output: Нет ✗

Russian Output with Ours: Да ✓

Figure 10: Case study.