CAAC: Confidence-Aware Attention Calibration to Reducing Hallucinations in Large Vision-Language Models

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) achieve impressive performance on multimodal tasks but often suffer from hallucination, and confidently describe objects or attributes not present in the image. Current inference-time interventions, while training-free, struggle to maintain accuracy in open-ended and long-form generation scenarios. We introduce the Confidence-Aware Attention Calibration (CAAC) framework to address this challenge by targeting two key biases: spatial perception bias, which distributes attention disproportionately across image tokens, and modality bias, which shifts focus from visual to textual inputs over time. CAAC employs a two-step approach: Visual-Token Calibration (VTC) to balance attention across visual tokens, and Adaptive Attention Re-Scaling (AAR) to reinforce visual grounding based on the model's confidence. This confidence-driven adjustment ensures consistent visual alignment during generation. Experiments on CHAIR, AMBER, and POPE benchmarks demonstrate that CAAC outperforms baselines, particularly in long-form generations, effectively reducing hallucination. Data and code are available at `https://anonymous.4open.science/r/CAAC-5D7F/`.

1 Introduction

Large vision-language models (LVLMs) (Bai et al., 2023; Chen et al., 2023; Liu et al., 2023; Chen et al., 2024; Dai et al., 2023; Ye et al., 2024) integrate visual and textual data using a pre-trained visual encoder, a cross-modal alignment module, and a powerful autoregressive decoder, enabling state-of-the-art performance in tasks such as image captioning, visual question answering, and visual reasoning. This multimodal capability has positioned LVLMs as key drivers in fields like content creation and human-computer interaction. However, a critical challenge is hallucination–generating content

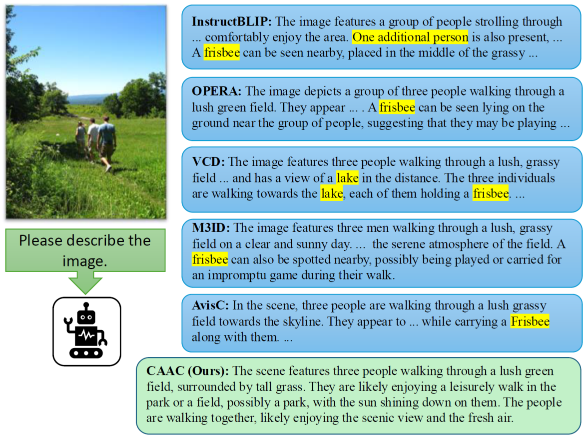


Figure 1: Comparison of the long-form generation (Max Generated Tokens: 512) of the baseline methods and our proposed CAAC framework. Hallucinations are highlighted in yellow.

ungrounded in the visual input, such as describing absent objects or misinterpreting scenes (Bai et al., 2025; Liu et al., 2024b; Li et al., 2023). This undermines the reliability of LVLMs, posing significant barriers to their deployment in safety-critical domains like medical diagnosis and autonomous navigation.

Efforts to mitigate hallucination in LVLMs have spawned a rich body of research, with strategies broadly classified into three categories: fine-tuning (Kim et al., 2023; Jiang et al., 2024; Gunjal et al., 2024), post-hoc rectification (Yin et al., 2023; Zhou et al., 2024), and inference-time interventions (Leng et al.; Huang et al.). Among them, inference-time interventions, due to their easy deployment and training-free nature, gained special momentum in the research community. Despite strong performance on discriminative tasks and short-form generation, existing methods struggle to maintain effectiveness in long-form generation. Figure 1 showcases an example of the failure of proposed hallucination mitigation methods under Max New Tokens of 512. This limitation stems from two fundamental mechanisms of LVLMs. First, spatial perception bias results in disproportionate attention to specific image regions, causing the model to overlook relevant visual cues. Second, modality bias causes the model to increasingly allocate more attention to textual information over visual input as generation progresses, leading to content that is poorly grounded in the image. Both biases significantly amplify the risk of hallucination in long-form generations.

To tackle these issues, we propose Confidence-Aware Attention Calibration (CAAC), a unified inference-time approach to mitigate hallucinations by dynamically recalibrating the LVLM's attention. CAAC uses the model's token-level confidence to adaptively adjust the attention distribution. Specifically, it counteracts both spatial perception bias and modality bias in a two-step process: an initial calibration smooths the attention maps of the decoder to prevent over-concentration on any single image region, and a subsequent confidence-guided reweighting increases the influence of the visual input whenever the chance of hallucination is high. By continuously reinforcing visual information when the model is uncertain, CAAC preserves visual grounding throughout the generation. As a result, CAAC effectively curbs hallucinations, even in challenging open-ended and long-form generation tasks, without sacrificing the fluency or detail of the generated text.

Our main contributions are summarized: (1) **Hallucination Analysis:** We present a novel analysis of hallucination in LVLMs using relevancy maps, which reveals two root causes of ungrounded generation. (2) **Mitigation Method:** We propose CAAC, an inference-time attention calibration framework, that adaptively calibrates the model's attention to promote visual grounding. (3) **Performance Improvement:** We demonstrate that CAAC significantly reduces hallucinations on multiple benchmarks for open-ended image captioning. In particular, our method outperforms state-of-the-art baselines, achieving an average 4% and 1.8% reduction in the hallucination rate compared with the best baseline on the CHAIR and AMBER benchmarks, respectively. Code and data are available at `https://anonymous.4open.science/r/CAAC-5D7F/`.

2 Related Work

A more detailed discussion of the related works is provided in Appendix C.

Large vision-language models (LVLMs) com-

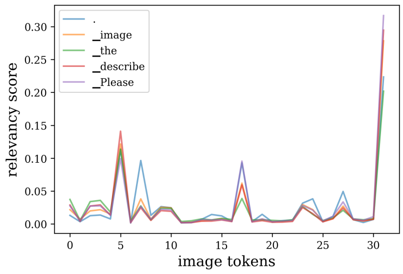


Figure 2: Distribution of image-token relevancy scores for InstructBLIP given a black canvas and the query "Please describe the image.". A pronounced skew toward a few image tokens can be witnessed.

bine visual encoders like CLIP (Radford et al., 2021) and ViT (Fang et al., 2023), cross-modal alignment modules such as linear projections (Liu et al., 2023) or Q-formers (Dai et al., 2023; Zhu et al., 2023), and language decoders like LLaMA (Touvron et al., 2023) or Vicuna (Zheng et al., 2023) to facilitate multimodal understanding and generation. State-of-the-art models, including mPLUG-Owl2 (Ye et al., 2024), InternVL (Chen et al., 2024), and QwenVL (Bai et al., 2023), utilize optimized architectures and diverse datasets to achieve strong performance in tasks like image captioning and visual reasoning (Xu et al., 2025).

Hallucination in LVLMs occurs when generated outputs do not accurately reflect visual inputs, posing challenges to their reliability (Guan et al., 2024; Liu et al., 2024b; Bai et al., 2025). Proposed mitigation strategies include fine-tuning techniques (Kim et al., 2023; Jiang et al., 2024; Liu et al., 2024a; Gunjal et al., 2024), post-hoc rectification methods (Yin et al., 2023; Zhou et al., 2024), and inference-time interventions (Leng et al.; Huang et al.; Woo et al., 2024; Suo et al., 2025; Favero et al.). Attention calibration, a training-free approach, has emerged as a promising solution to reduce hallucinations (Zhu et al., 2025; Zhang et al., 2024; Liu et al., 2024c; Gong et al., 2024; Woo et al., 2024). Our method builds on the insights derived from the previous works but introduces an adaptive intervention based on the model's confidence in predicting the next token.

3 Proposed Method

What causes LVLMs to describe objects or scenes absent from an image confidently? Our analysis identifies two primary culprits: spatial perception bias (Zhu et al., 2025), a skewed attention distribution favoring specific image tokens regardless of content, and modality bias, an increasing reliance on language priors over visual inputs as generation progresses. To tackle these challenges, we propose the Confidence-Aware Attention Calibration (CAAC) framework, which integrates two steps: an initial Visual-Token Calibration (VTC) to mitigate spatial perception bias by smoothing attention spikes across image tokens, and a confidence-driven Adaptive Attention Re-Scaling (AAR) to counteract modality bias by enhancing visual grounding throughout generation. Next, we detail the inference process in LVLMs (Sec. 3.1), present our novel analysis using relevancy maps to uncover these biases (Sec. 3.2, 3.3), and introduce CAAC's components (Sec. 3.4), highlighting their synergistic design to improve reliability.

3.1 Inference in LVLMs

Large vision–language models generate text conditioned on both an input image and a text prompt. An image is first encoded into visual tokens via a pre-trained vision encoder. The visual tokens are then mapped into the language embedding space using a linear projection or a more complex alignment module to extract textual information from the image, yielding image tokens $I = \{i_1, \dots, i_{N_i}\}$. Concurrently, the text query is also tokenized into N_q tokens $Q = \{q_1, \dots, q_{N_q}\}$. Then, the LLM decoder parameterized by θ receives concatenated embeddings (I, Q) and auto-regressively generates a sequence of N_g tokens $G = \{y_1, \dots, y_{N_g}\}$. Formally, at t'th generation round, the next token is drawn from the following probability distribution:

$$y_t \sim p_\theta(y_t | I, Q, y_{<t}) \tag{1}$$

where $y_{<t} = \{y_1, \dots, y_{t-1}\}$ is the sequence of previously generated tokens. Various sampling strategies have been developed for efficient and controllable sampling from the probability distribution (Shi et al., 2024). The generation process continues until the End-of-Sequence (EOS) token is selected or the maximum allowed number of tokens is reached.

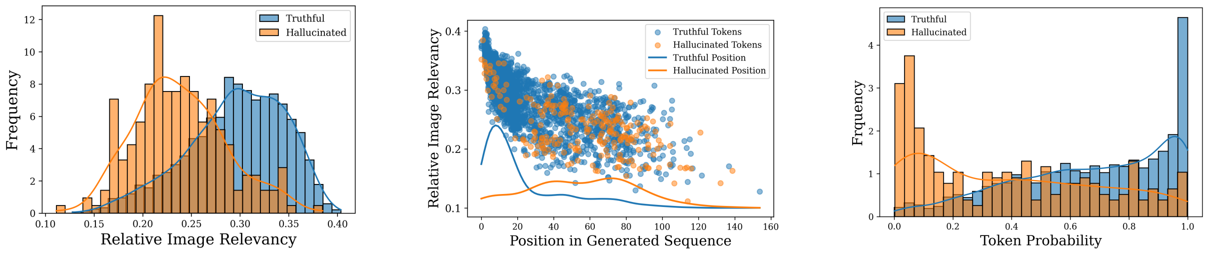
3.2 Analysis: Disproportionate attention to image tokens

Previous studies have shown that LVLM decoders tend to concentrate attention on a small subset of visual tokens—termed attention sinks(Zhang et al., 2024), summary tokens(Huang et al.), or blind tokens(Woo et al., 2024)—regardless of image content, including blank inputs. This phenomenon, also known as spatial perception bias(Zhu et al., 2025), has been linked to downstream hallucination errors (Huang et al.; Zhang et al., 2024). While our analysis is motivated by similar concerns, we identify a key methodological limitation in prior work: **their conclusions are based on raw attention weights from individual layers**, which do not reliably reflect token importance. Indeed, token embeddings are progressively contextualized across layers, meaning that accurate attribution requires tracing the influence of each input token through the entire network.

To address this limitation, we leverage **relevancy maps** (Chefer et al., 2021), which propagate token-level contributions layer by layer, ultimately quantifying the influence of each input token on the generation of each output token. By adopting this more principled analysis, our work revisits and reinterprets previous findings, offering new insights. We observe that given a black canvas image and a standard query, less than 10% of image tokens accumulate more than 50% of relevancy scores, while the vast majority of image tokens contribute minimally (Figure 2). This distribution remains consistent across various meaningless inputs and queries (see Appendix B), underscoring a robust bias pattern: **The decoder assigns disproportionate attention to image tokens, leading to the model's over-reliance on a few image tokens, thereby increasing the likelihood of hallucination.**

3.3 Analysis: Decaying attention to image tokens

Another significant contributor to LVLM hallucination is the model's increasing reliance on its text history at the expense of visual inputs, particularly in open-ended tasks like image captioning. Prior work has shown that when the model is uncertain, language priors often dominate the generation process (Zhou et al., 2024). To quantify this, we leverage AMBER's generative pipeline, prompting InstructBLIP (Dai et al., 2023) to describe each image in detail. Then, we extract truthful and hallucinatory tokens using predefined hallucinatory and truthful object sets from AMBER. We compute the *relative image relevancy* by the relevancy map framework to quantify the aggregate contribution of all image tokens to the generation of each output token. For an input comprising I image tokens and T text tokens (total $N = I + T$), the relative image



(a) (b) (c)

Figure 3: (a) Normalized histogram of relative image relevancy scores for truthful (blue) and hallucinatory (orange) tokens, showing higher image relevancy for truthful tokens. (b) Scatter plot of relative image relevancy versus absolute position in the generated sequence. Every point represents one generated token (truthful or hallucinatory), and the lines indicate the density of token positions. (c) Normalized histogram of logit probabilities for truthful vs. hallucinatory tokens, showing lower probabilities for hallucinatory tokens. Best viewed in color.

relevancy at generation step t is defined as:

$$R_{rel_N} = \frac{\sum_{i=1}^{I} R^{iN}}{\sum_{j=1}^{N} R^{jN}} \tag{2}$$

where R^{ij} represents the influence of i'th token on j'th token. Figure 3a shows the distribution of relative image relevancy for truthful and hallucinatory tokens. There is a statistically significant difference between the two distributions, suggesting that hallucinatory tokens have markedly lower relative image relevancy. Moreover, relative image relevancy declines as the generated sequence lengthens (Figure 3b). This decay confirms that extended generation increases the model's tendency to overlook visual inputs, a phenomenon we term modality bias, reflecting a preference for textual over visual information. The other takeaway from Figure 3b is that the hallucinatory tokens appear later in the generated sequence, underscoring the importance of mitigating hallucinations in long-form generations.

We also examine the generation confidence by inspecting token logit probabilities (Figure 3c). We find that truthful tokens are heavily skewed toward high probabilities, whereas hallucinatory tokens are skewed toward the low-probability regime. It suggests a distinct generation dynamic between truthful and hallucinatory tokens: **the model hallucinates when its confidence is low and its attention to the image has diminished**.

3.4 CAAC Framework

Our CAAC framework addresses two distinct biases operating in different dimensions within the LLM decoder. Spatial perception bias is a universal, query-agnostic distortion in attention distribution across image tokens. In contrast, modality bias operates at the token level, increasingly skewing attention toward textual inputs as generation length extends. CAAC tackles these challenges through a unified attention calibration strategy, featuring two components: Visual-Token Calibration (VTC), which corrects the universal spatial perception bias by adjusting attention weights, and Image Attention Upscaling (IAU), which mitigates modality bias by adaptively amplifying visual information during the generation. This integrated approach ensures a balanced multimodal processing, enhancing LVLM reliability.

3.4.1 Visual-Token Calibration (VTC)

The VTC module aims to mitigate spatial perception biases in LVLMs by adjusting the attention distribution over image tokens within the decoder's attention heads. By targeting the attention from the final query token to image tokens and applying a calibration derived from a reference input, we achieve a more balanced attention distribution while preserving essential visual information.

In LVLMs, the attention mechanism of the decoder plays a pivotal role in integrating visual and textual information. Specifically, the attention from the last query token to image tokens directly informs the prediction of the subsequent token, making it a critical point of intervention. Given an input comprising visual tokens $I = \{i_1, i_2, \ldots, i_{N_i}\}$ and query tokens $Q = \{q_1, q_2, \ldots, q_{N_q}\}$ ($N = I + Q$), the attention map for a given head h in layer l is denoted $A^{h,l} \in \mathbb{R}^{(N_i+N_q)\times(N_i+N_q)}$. We focus on the submatrix corresponding to the last query token's attention to image tokens, i.e., the last row's first N_i columns, defined as $V^{h,l} = [A^{h,l}_{N,j}]_{j\in\mathcal{I}} \in \mathbb{R}^{N_i}$.

Calibration Vector Construction: To establish a baseline for calibration, we use a reference input consisting of a meaningless image and a generic

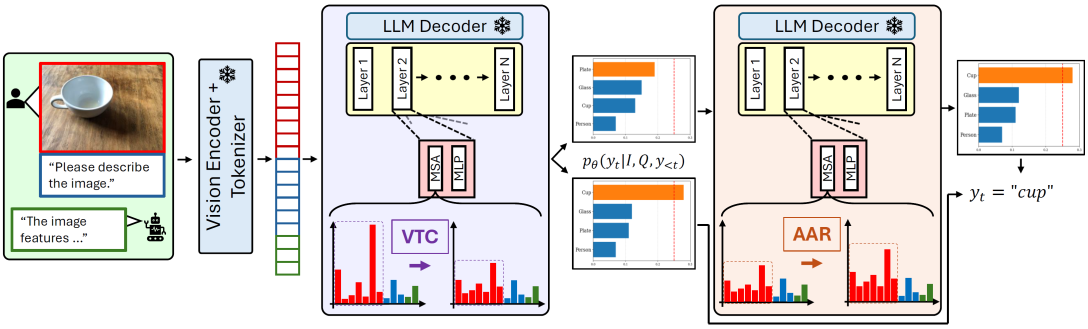


Figure 4: Overview of the CAAC Framework. The CAAC framework comprises two key components: VTC, which adjusts skewed attention to image tokens to reduce spatial perception bias, and AAR, which adaptively augments attention to image tokens to address modality bias. Both components are applied to the multi-head self-attention (MSA) module within the decoder.

query (e.g., "What is this?"). Choosing a meaningless image ensures that attention patterns reflect the model's baseline behavior rather than meaningful content, and empirical tests show that the choice of the meaningless image has no meaningful impact on the resulting calibration (Appendix B). For each attention head h in layer l, we extract $V^{h,l}$ from the reference input's attention map. Alternatively, to enhance robustness, $V^{h,l}$ may be computed as the average of the last few rows' image-token columns.

Therefore, given the vector $V^{h,l} \in \mathbb{R}^{N_i}$, where $V^{h,l} = [v_1, v_2, \dots, v_{N_i}]$ and $v_i \neq 0$ for all i, the initial inverse is computed as:

$$V^{h,l}_{\text{cal},0} = [1/v_1, 1/v_2, \dots, 1/v_{N_i}] \tag{3}$$

To ensure the sum of entries remains consistent with the original vector, we scale $V^{h,l}_{\text{cal},0}$ by the ratio of the sum of $V^{h,l}$ to the sum of $V^{h,l}_{\text{cal},0}$. The final calibration vector is thus:

$$V^{h,l}_{\text{cal}} = \frac{\sum_{i=1}^{N_i} v_i}{\sum_{i=1}^{N_i} (1/v_i)} \cdot V^{h,l}_{\text{cal},0}, \tag{4}$$

where $\sum_{i=1}^{N_i} v_i$ is the sum of the original attention weights, and $\sum_{i=1}^{N_i} (1/v_i)$ is the sum of the initial inverted weights. Note that the product of $V^{h,l}$ and $V^{h,l}_{\text{cal}}$ results is a uniform vector with the same sum as $V^{h,l}$. This inversion counteracts the skew attention pattern of the image tokens.

Application of Calibration: For a specific input image and query pair, let $V \in \mathbb{R}^{N_i}$ represent the attention from the last query token to image tokens in the attention map $A^{h,l}$. We flatten this by computing the element-wise product $V_u = V \odot V^{h,l}_{\text{cal}}$, where $\odot$ denotes the Hadamard product. V_u approximates a uniform attention distribution across image tokens. However, enforcing strict uniformity can distort visual information, as positional embeddings naturally differentiate image token representations, even for identical patches. This differentiation is naturally reflected in the attention scores received by different image tokens.

Smoothing with Parameter β: To balance bias correction and information preservation, we introduce a smoothing parameter $\beta \in [0, 1]$ to control smoothing. The smoothed attention vector V_s is computed as a weighted average of the original and calibrated vectors:

$$V_s = (1 - \beta)V + \beta V_u \tag{5}$$

When $\beta = 0$, the original attention V is retained and when $\beta = 1$, the fully calibrated V_u is applied, yielding a near-uniform distribution. Intermediate values of β allow for promoting more balanced attention distribution without over-correcting the attention distribution. This flexibility ensures that the calibration enhances model reliability and is what makes the VTC module different than UAC (Zhu et al., 2025).

3.4.2 Adaptive Attention Re-Scaling (AAR)

The Adaptive Attention Re-Scaling (AAR) module is designed to mitigate modality bias, where attention to image tokens diminishes over time during autoregressive generation. AAR counteracts this by dynamically increasing the attention from the last query token to image tokens, reinforcing visual grounding throughout the generation sequence, particularly when the model's predictions falter. AAR focuses on the same segment of the attention map as the VTC module, specifically the attention vector $V^{h,l} = [A^{h,l}_{N,j}]_{j \in \mathcal{I}} \in \mathbb{R}^{N_i}$ to steer model's attention toward visual information by scaling up the attention weights of visual tokens.

Confidence-Aware Scaling: AAR operates autoregressively, adjusting attention in every generation round to maintain visual relevance across the entire sequence. A key question is: *what is the appropriate scaling factor, as token dependency on visual input varies?* Tokens essential for text cohesion (e.g., conjunctions) require minimal intervention, whereas image-dependent tokens (e.g., nouns and adjectives describing visual content) demand stronger visual grounding. Our analysis revealed that hallucinatory tokens often emerge when the model lacks confidence (Figure 3c). This insight drives AAR's adaptive strategy: **scaling is triggered by the model's uncertainty**.

In generation round t, a forward pass computes the maximum logit probability p_t for the predicted token.

$$p_t = \max_y p_\theta(y \mid I, Q, y_{<t}). \tag{6}$$

If p_t falls below a preset threshold p_{thr}, AAR calculates a scaling factor λ as a probability-weighted average of set minimum and maximum scale factor:

$$\lambda_t = \lambda_{\text{min}} \cdot p + \lambda_{\text{max}} \cdot (1 - p) \tag{7}$$

With $\lambda_{\text{min}} = 1$ we ensure no scaling is applied when the model is fully confident ($p = 1$), while λ_{max} sets the upper bound for scaling when confidence is minimal ($p = 0$). As p decreases, λ increases, amplifying attention to image tokens precisely when hallucination risk is highest.

Application of AAR: As AAR is bound to change the sum of the row it is applied to, we need to apply it to the attention weights before softmax. After the intervention, softmax is applied normally to ensure all rows sum to 1. When $p < p_{\text{thr}}$, the attention vector before softmax $V^{h,l}$ is scaled:

$$V^{h,l}_{\text{t, scaled}} = \lambda_t \cdot V^{h,l}_t \tag{8}$$

This scaled vector replaces the original vector in the decoder's attention mechanism, shifting focus toward visual inputs. If $p \geq p_{\text{thr}}$, no scaling occurs, preserving the model's natural behavior.

4 Experimental Results

4.1 Setup

Models. We evaluate our method on two widely adopted open-source LVLMs: InstructBLIP and LLaVA-1.5, both configured with 7 billion parameters. We particularly select these models for direct comparison with existing baselines (Leng et al.; Huang et al.; Favero et al.); however, our CAAC framework is model-agnostic and can seamlessly integrate with any LVLM.

Benchmarks To evaluate the effectiveness of CAAC in reducing hallucinations in long-form generations, we prioritize generative benchmarks that support open-ended outputs. We adopt CHAIR (Rohrbach et al., 2019) and AMBER (Wang et al., 2024) as our generative benchmarks, alongside POPE MSCOCO (Li et al., 2023) as the discriminative benchmark to provide a comprehensive evaluation of CAAC.

Metrics. We prioritize metrics that directly measure hallucination rates, such as CHAIR_i and CHAIR_s for the CHAIR benchmark, and CHAIR and HAL for the AMBER benchmark, due to their critical role in assessing the model's factual alignment with visual input. While we report Recall scores for CHAIR and COVER scores for AMBER, which evaluate the informativeness and exhaustiveness of generated responses, these metrics are less relevant to our primary objective. High Recall or COVER scores paired with elevated hallucination rates can lead to misleading outputs, as the model may generate exhaustive but factually incorrect descriptions, undermining reliability. For the POPE benchmark, we report Accuracy and F1 scores to complement our evaluation.

Baselines. We compare against four inference-time mitigation methods that require no additional training. Contrastive decoding methods include VCD (Leng et al.), AvisC (Woo et al., 2024), and M3ID (Favero et al.), which mitigate hallucinations via a contrastive decoding technique, and OPERA (Huang et al.), a beam-search modification that penalizes over-trusted tokens to promote visual grounding.

Implementation Details. For the baselines, we adopt the hyperparameter settings reported in their respective papers to ensure consistency. For CAAC, we set the smoothing parameter $\beta = 0.7$ for LLaVA and $\beta = 0.5$ for InstructBLIP. The maximum scaling factor for AAR is set to $\lambda_{\text{max}} = 1.5$ for both tasks, with $\lambda_{\text{min}} = 1.0$ and $p_{thr} = 0.25$. More experimental details are presented in Appendix A.

4.2 Comparison to Baselines

CHAIR. The CHAIR benchmark (Rohrbach et al., 2019) evaluates object hallucination in image captioning by measuring, for a given input image and a corresponding caption, the fraction of hallucinated objects, CHAIR_i, and the the fraction

Table 1: Performance on CHAIR Benchmark

Model	CHAIRs	CHAIRi	Recall	Len
LLaVA	55.2	17.6	73.8	103.9
+ OPERA	<u>44.6</u>	**12.8**	79.2	
+ VCD	57.8	16.3	78.3	103.4
+ AvisC	60.4	17.2	78	104
+ M3ID	56.2	16.4	81.1	93.7
+ CAAC	**40.8**	<u>13</u>	75.5	85.9
InstructBLIP	55.6	16.6	71.1	111.2
+ OPERA	<u>46.4</u>	<u>14.2</u>	72.9	92.6
+ VCD	60.8	17.9	73.7	107.1
+ AvisC	71	20.1	71.4	98.9
+ M3ID	72.8	21.1	71.7	103.1
+ CAAC	**37.4**	**10.8**	72.6	88.4

of hallucinated sentences, CHAIR_s. We used the same evaluation setting as OPERA. We also used the same subset of 500 images from the validation set of the COCO 2014 dataset (Lin et al., 2015), paired them with the prompt "Please describe this image in detail.", and collected the responses from LVLM. We set Max Tokens to 512 to avoid prematurely truncating generation sequences. Table 1 summarizes the results of the CAAC framework and the baselines on the CHAIR benchmark. As shown, CAAC effectively reduces the hallucination rates in both CHAIR_i and CHAIR_s while maintaining a comparable recall score with most baseline methods.

AMBER. The AMBER benchmark (Wang et al., 2024) assesses hallucinations in LVLMs through generative and discriminative tasks, focusing on object existence, attributes, and relationships. We concentrate on the generative task, conducting experiments under two settings: Max Tokens 64, aligning with baseline configurations, and Max Tokens 512, to evaluate performance with longer sequences. AMBER employs several metrics to evaluate the generated text, including CHAIR, Hal (the proportion of responses with hallucinations), Cover (the proportion of image objects mentioned in the text), and Cog (the proportion of hallucinations aligned with human cognition/expectation).

Our CAAC framework excels on the AMBER benchmark, delivering the lowest hallucination rates in CHAIR and HAL metrics for the Max Tokens 512 setting (Table 2). For the Max Tokens 64 setting with LLaVA, OPERA performs similarly on hallucination metrics, with slightly higher CHAIR and lower HAL and COG values. Contrastive decoding techniques, however, show significant degradation in managing hallucinations during long generations, underscoring their limitations. CAAC effectively reduces hallucinations across both short and long sequences, achieving coverage scores on par with baselines, notably matching OPERA—the closest competitor in hallucination metrics, while surpassing the base model's recall, thus improving the accuracy and informativeness.

POPE. The Polling-based Object Probing Evaluation (POPE) benchmark (Li et al., 2023) provides a streamlined approach to assess object hallucination in Large Vision-Language Models by querying whether specific objects exist in a given image. POPE employs three sampling settings for negative samples—random, popular, and adversarial—each designed to challenge the model's discriminative capabilities differently. Although our CAAC framework is primarily designed for generative tasks, it exhibits robust performance in this discriminative setting, as shown in Table 3. CAAC achieves competitive Accuracy and F1 scores across all settings and for both LLaVA and InstructBLIP. These results highlight CAAC's effectiveness in mitigating hallucinations beyond its generative focus, outperforming or matching baseline methods, thus demonstrating its versatility and robustness.

4.3 Ablation Study

To measure the influence of each module within the CAAC framework, we conducted ablation experiments using the InstructBLIP-7B model on the AMBER and CHAIR benchmarks. We evaluated four configurations: the baseline model, VTC-only, AAR-only, and the full CAAC framework incorporating both modules. These experiments were performed under the same settings as those used for the generative tasks, as outlined in Section 4.1.

The results are presented in Figure 6, which includes the results for the CHAIR benchmark and the AMBER benchmark. As shown, both modules individually contribute to lowering hallucination rates, as measured by CHAIR and Hal metrics, while also increasing recall. Also, the full CAAC framework achieves the most significant improvements across all metrics, guiding the model toward more informative and accurate generations.

4.4 Hyperparameter Analysis

We optimized the CAAC framework by tuning its key parameters, focusing on the Adaptive Attention Re-Scaling (AAR) and Visual-Token Calibration (VTC) modules to balance hallucination reduction

Table 2: Performance on AMBER Benchmark Across Different MaxTokens Settings

Model	MaxTokens 64 CHAIR↓	HAL↓	COG↓	COVER↑	MaxTokens 512 CHAIR↓	HAL↓	COG↓	COVER↑
LLaVA	7.95	31.0	2.2	44.5	11.3	48.1	4.3	50.4
+ OPERA	5.10	**19.1**	**1.5**	45.00	7.3	29.5	3.1	47.5
+ VCD	6.70	27.8	1.95	46.50	8.2	37.3	3.9	51.9
+ M3ID	6.00	26.0	1.5	48.90	7.2	41.4	3.1	57.3
+ AvisC	6.25	25.6	2	46.55	11.0	48.0	5	52.5
+ CAAC (Ours)	**4.90**	19.7	1.9	45.40	**6.0**	**24.8**	**2.5**	47.6
InstructBLIP	9.6	36	2.3	46.5	12.8	53.5	5.2	52.7
+ OPERA	6.60	24.7	2.5	46.40	9.7	40.5	4.5	51.2
+ VCD	7.60	29.9	2.3	47.65	10.8	46.6	4.9	53.4
+ M3ID	6.90	27.5	2.2	47.20	10.4	47.3	4.5	51.7
+ AvisC	6.70	28.0	2.5	46.65	10.1	46.8	4.5	51.2
+ CAAC (Ours)	**4.8**	**20.3**	**2**	46	**5.6**	**25.8**	**2.6**	47.8

Table 3: Performance on POPE MSCOCO Benchmark Across Different Sampling Settings

Model	Random Accuracy	Random F1	Popular Accuracy	Popular F1	Adversarial Accuracy	Adversarial F1
LLaVA	83.77	81.94	82.57	80.86	79.77	78.47
+ OPERA	**88.49**	**88.48**	83.40	**86.4**	81.20	**82.24**
+ VCD	85.43	83.99	83.17	81.94	80.27	79.49
+ M3ID	86.13	81.85	82.07	80.77	79.5	78.15
+ AvisC	84.67	82.21	83.67	81.27	**81.83**	79.55
+ CAAC (Ours)	88.47	87.8	**85.93**	85.5	81.03	81.37
InstructBLIP	81.53	81.19	78.47	78.75	77.43	78
+ OPERA	**89.18**	**88.68**	**83.97**	**83.69**	**81.83**	**81.91**
+ VCD	82.03	81.56	79.13	79.2	77.23	77.72
+ M3ID	82.33	81.53	80.9	80.42	78.53	78.49
+ AvisC	86.03	84.41	84.27	82.77	81.83	80.67
+ CAAC (Ours)	87.67	87.05	83.47	83.38	81.17	81.53

while preserving response quality. For AAR, we set the confidence threshold $p_{\text{thr}} = 0.25$, $\lambda_{\max} = 1.5$, and applied it to all decoding layers, achieving consistent and coherent outputs. For VTC, applying it to the first 10 layers (out of 32) minimized hallucination rates effectively, avoiding the incoherence or truncated sequences observed with full-layer application. The smoothing parameter β was found to be very impactful. Large values of β (≥ 0.9) often resulted in impaired generation sequences. However, intermediate values for β, $0.3 \sim 0.7$, resulted in coherent and high-quality responses. A comprehensive analysis of the models' settings is provided in Appendix D.

4.5 Qualitative Evaluation

To assess the qualitative performance of our CAAC framework, we compare its outputs with those of baseline methods on the AMBER dataset. For instance, Figure 1 shows the captions generated via InstructBILP and those of the baselines for a given image. The hallucinations are highlighted. Despite the baseline methods, CAAC accurately suppresses the word frisbee, which is not present in the image. Additional examples are provided in the Appendix E.

5 Conclusion

We introduced the Confidence-Aware Attention Calibration (CAAC), a training-free, inference-time framework that mitigates hallucination in LVLMs by addressing spatial and modality biases through Visual-Token Calibration and Adaptive Attention Re-Scaling, ensuring consistent visual grounding across diverse generation tasks. Experiments on benchmarks like CHAIR, AMBER, and POPE MSCOCO demonstrate CAAC's effectiveness in reducing hallucination rates, surpassing baselines like OPERA, particularly in long sequences, despite a trade-off with metrics like COVER and Recall. This prioritization of factual accuracy over exhaustive detail makes CAAC a practical solution for enhancing LVLM reliability in safety-critical applications.

6 Limitations

Extra Inference Time. One limitation of our CAAC framework is the potential increase in inference time due to the need for two forward passes for certain tokens. The Adaptive Attention Re-Scaling (AAR) module requires an initial forward pass to compute logit probabilities, and if the maximum logit probability falls below the preset threshold (e.g., 0.25), a second pass is needed with adjusted attention weights. Our experiments on the CHAIR benchmark, generating detailed descriptions for 500 MS COCO images with 512 max new tokens and a probability threshold of 0.25, revealed that only 14% of tokens required a second pass, indicating a modest impact on inference time. This trade-off is acceptable in factually critical domains like healthcare, where the reduction in hallucinations outweighs the slight latency, and the variability (dependent on threshold, sequence length, and input) is a reasonable scope limitation given our focus on accuracy over speed.

Suboptimal recall scores. Another limitation is that CAAC may compromise recall scores while mitigating hallucinations, as it steers the model's attention toward visual information during long generations, preventing deviations from the input image and lowering the generation length by a few percent (Table 1). This intentional focus could lessen the model's ability to generate exhaustive descriptions. In our experiments, recall sometimes dipped relative to the strongest hallucination baselines, though it still exceeded that of the base model. This trade-off aligns with our goal of enhancing reliability in safety-critical applications, where factual correctness is paramount.

Model-specific hyper-parameter tuning. A third limitation is the need to tune parameters such as β, $\lambda_{\max}$, and the number of decoder layers for VTC intervention independently for each LVLM, adding an overhead step before deployment. Yet, this approach remains computationally efficient compared to methods requiring training post-hoc hallucination correction modules or fine-tuning the entire model, a common practice in baseline approaches. Given that our work focuses on inference-time intervention rather than training, this tuning requirement is a reasonable trade-off and does not detract from the effectiveness of CAAC as a practical solution for hallucination mitigation.

Ethical Considerations

Our research complies with ethical standards, using publicly available datasets like MS COCO, AMBER, and CHAIR, sourced responsibly under their licenses. The content of these datasets does not reflect the authors' views, and no personally identifiable information was involved. We ensure transparent reporting to support fairness and reproducibility.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint*. ArXiv:2308.12966 [cs].

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of Multimodal Large Language Models: A Survey. *arXiv preprint*. ArXiv:2404.18930 [cs].

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396, Montreal, QC, Canada. IEEE.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint*. ArXiv:2306.15195 [cs].

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint*. ArXiv:2312.14238 [cs].

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint*. ArXiv:2305.06500.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, Vancouver, BC, Canada. IEEE.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro

Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-Modal Hallucination Control by Visual Information Grounding.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. DAMRO: Dive into the Attention Mechanism of LVLM to Reduce Object Hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7696–7712, Miami, Florida, USA. Association for Computational Linguistics.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. *arXiv preprint*. ArXiv:2310.14566 [cs].

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143. Issue: 16.

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.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Jae Myung Kim, A. Koepke, Cordelia Schmid, and Zeynep Akata. 2023. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2585–2595.

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. *arXiv preprint*. ArXiv:2305.10355 [cs].

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. *arXiv preprint*. ArXiv:1405.0312 [cs].

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. *arXiv preprint*. ArXiv:2306.14565 [cs].

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A Survey on Hallucination in Large Vision-Language Models. *arXiv preprint*. ArXiv:2402.00253 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *arXiv preprint*. ArXiv:2304.08485 [cs].

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. *arXiv preprint*. ArXiv:2407.21771 [cs].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*. ArXiv:2103.00020 [cs].

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object Hallucination in Image Captioning. *arXiv preprint*. ArXiv:1809.02156 [cs].

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A Thorough Examination of Decoding Methods in the Era of LLMs. *arXiv preprint*. ArXiv:2402.06925 [cs].

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating Hallucination via Dynamic Contrastive Decoding. *arXiv preprint*. ArXiv:2503.00361 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971 [cs].

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. *arXiv preprint*. ArXiv:2311.07397 [cs].

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2024. Don't Miss the Forest for the Trees: Attentional Vision Calibration for Large Vision Language Models. *arXiv preprint*. ArXiv:2405.17820 [cs].

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2025. LVLM-EHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1877–1893.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv preprint*. ArXiv:2304.14178 [cs].

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *arXiv preprint*. ArXiv:2310.16045 [cs].

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024. Seeing Clearly by Layer Two: Enhancing Attention Heads to Alleviate Hallucination in LVLMs. *arXiv preprint*. ArXiv:2411.09968 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint*. ArXiv:2306.05685 [cs].

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. *arXiv preprint*. ArXiv:2310.00754 [cs].

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint*. ArXiv:2304.10592 [cs].

Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. 2025. Mitigating Object Hallucinations in Large Vision-Language Models via Attention Calibration. *arXiv preprint*. ArXiv:2502.01969 [cs].

A Further Experimental Details

Baselines. We compare CAAC with four training–free, inference-time approaches:

- **VCD** (Visual Contrastive Decoding) (Leng et al.): each token is decoded twice—once with the original image and once with a perturbed copy—and words whose likelihood collapses under perturbation are down-weighted.
- **AvisC** (Woo et al., 2024): first performs *attentional vision calibration* by masking high-attention outlier tokens, then applies contrastive decoding to suppress visually ungrounded candidates.
- **M3ID** (Favero et al.): re-scores candidate tokens with a lightweight image-guided gradient signal, promoting those whose gradients align with visual features and filtering hallucinations.
- **OPERA** (Huang et al.): augments beam search with an *over-trust penalty* and a *retrospection–allocation* term, penalising tokens that receive insufficient cumulative attention from the image.

Implementation Details.

Hardware and runtime. All experiments were run on a single server equipped with 4×H100 40 GB GPUs and 512 GB of system RAM. We evaluate models in 16-bit floating-point precision using HuggingFace `transformers 4.47`. A complete AMBER run (512 max-token setting) requires ∼12 hours for InstructBLIP and ∼10 hours for LLaVA-1.5.

CAAC hyper-parameters. We use the following values, selected via the grid search: $\beta = 0.7$ for LLaVA-1.5 and $\beta = 0.5$ for InstructBLIP; $\lambda_{\text{max}} = 1.5$, $\lambda_{\text{min}} = 1.0$; confidence threshold $p_{\text{thr}} = 0.25$; VTC applied to the first 10 decoder layers (out of 32). For discriminative task (POPE) with InstructBLIP, we set $\lambda_{min} = 0$, $lambda_{max} = 1.8$, and applied VTC on the first five layers of the decoder. This setting optimized the performance of InstructBLIP for discriminative tasks while ensuring that the scale factor is greater than 1 when $p < P_{thr}$.

Baseline implementations and settings.

- **OPERA** (Huang et al.): official code[1] with `beam_size=5`, `num_cands=5`, scale factor 50, $\alpha = 1, \beta = 5, r = 15$.
- **Contrastive decoding baselines.** We adopt the official `AvisC` repository[2] for all three CD variants and keep the authors' recommended hyper-parameters:
  - **VCD** (Leng et al.): $\alpha = 1$, $\beta = 0.1$, $\gamma = 0.1$.

[1] https://github.com/shikiw/OPERA
[2] https://github.com/sangminwoo/AvisC

- **AvisC** (Woo et al., 2024): $\gamma = 0.5$, $\lambda = 1$, $\alpha = 2.5$ (LLaVA) / 3.0 (InstructBLIP).
- **M3ID** (Favero et al.): $\lambda = 0.2$ (default).

For all baselines, we retain the original decoding parameters (temperature, top-p, etc.) reported in their papers to ensure fair comparison with our CAAC framework.

B Image Attention Skew

Visual-Token Calibration (VTC) relies on a single "reference" image to derive its calibration vector. A natural concern is whether the choice of that reference—white canvas, black canvas, or random noise—affects the resulting adjustment. To test this, we feed each meaningless image to the LVLM together with the fixed query *"Please describe the image."* and compute the relative image-relevancy for all query tokens. Figure 5 shows the results for InstructBLIP and LLaVA. As one can see from the relative image relevancy plots, the choice of the reference image for calibration has no meaningful impact on the calibration vectors.

C Related Work

C.1 Large Vision-Language Models

Large vision–language models (LVLMs) bring together powerful visual backbones and large language models to enable rich multimodal understanding and generation. At their core, LVLMs consist of three components: a pretrained visual encoder (e.g., CLIP (Radford et al., 2021), ViT (Fang et al., 2023)) that extracts image embeddings; a lightweight cross-modal alignment module, ranging from a simple linear projection (Liu et al., 2023) to more sophisticated "Q-former" architectures (Dai et al., 2023; Zhu et al., 2023), that maps these visual features into the LLM's embedding space; and a large autoregressive language decoder (e.g., LLaMA (Touvron et al., 2023), Vicuna (Zheng et al., 2023)) that generates fluent text. Increasingly advanced LVLM families, such as mPLUG-Owl2 (Ye et al., 2024), InternVL (Chen et al., 2024), and QwenVL (Bai et al., 2023), have also been proposed, driven by diverse data, optimized architectures, and training paradigms. LVLMs, thanks to their unified pipeline, acheived state-of-the-art results on tasks such as open-ended image captioning, visual question answering, visual reasoning, etc. (Xu et al., 2025)

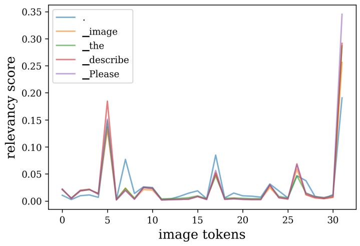
C.2 Hallucination in LVLMs

Hallucination in Large Vision-Language Models (LVLMs) refers to the generation of responses that are not factually aligned with the visual input, such as describing objects absent from the image or misinterpreting visual content (Guan et al., 2024; Liu et al., 2024b; Bai et al., 2025). This phenomenon poses a significant challenge to the reliability and practical deployment of LVLMs in real-world applications. To address this issue, the literature proposes several mitigation strategies, broadly categorized into fine-tuning approaches (Kim et al., 2023; Jiang et al., 2024; Liu et al., 2024a; Gunjal et al., 2024), post-hoc rectification techniques (Yin et al., 2023; Zhou et al., 2024), and inference-time interventions (Leng et al.; Huang et al.; Woo et al., 2024; Suo et al., 2025; Favero et al.). A promising direction within inference-time interventions is calibrating attention mechanisms within LVLMs. (Zhu et al., 2025) Introduces uniform and dynamic attention calibration to remove spatial perception bias present in LVLMs. (Zhang et al., 2024) mitigates hallucination by strengthening the influence of dense attention sinks in the early layers of the decoder. (Liu et al., 2024c) uses visual contrastive decoding while increasing the weight of the attention to image tokens in the self-attention heads of the decoder. (Gong et al., 2024) uses the CLS token from the vision encoder to filter out high-attention outlier tokens via a contrastive decoding strategy. (Woo et al., 2024) constructs the modified visual input by zeroing out the attention to all tokens except blind tokens and uses a contrastive decoding scheme to reduce hallucination.

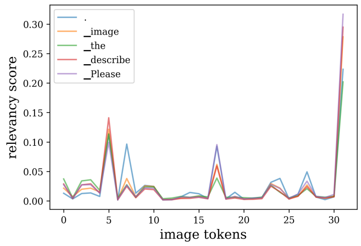
D Hyperparameter Analysis

In this subsection, we evaluate the impact of key parameters in the CAAC framework on its performance, focusing on the most influential ones due to limited computational resources. For the Adaptive Attention Re-Scaling (AAR) module, we set the confidence threshold $p_{\text{thr}} = 0.25$, as hallucinatory token frequency increases noticeably when the logit probability drops below this value (3c). We also selected $\lambda_{\max} = 1.5$, since values above 2 impair response fluency and coherence. Furthermore, applying AAR to all decoding layers proved optimal, yielding consistent and coherent outputs based on experimental results.

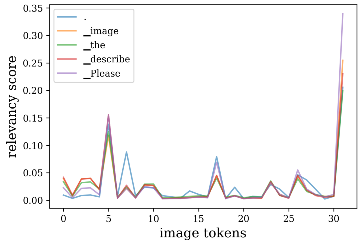
For the Visual-Token Calibration (VTC) module, applying it to all layers often produced incoherent



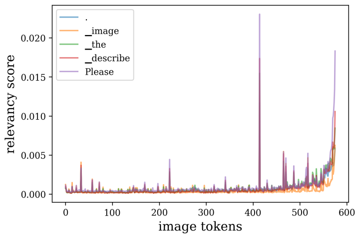
(a) Black image (InstructBLIP)



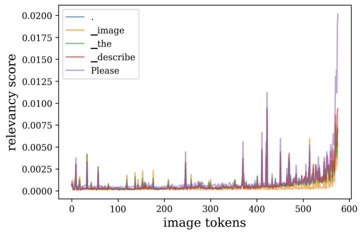
(b) White image (InstructBLIP)



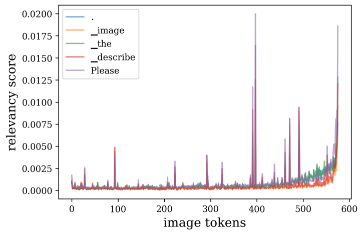
(c) Noise image (InstructBLIP)



(d) Black image (LLaVA)

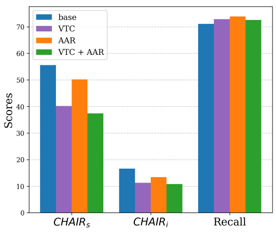


(e) White image (LLaVA)

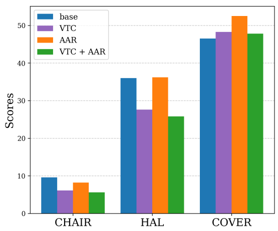


(f) Noise image (LLaVA)

Figure 5: Distribution of relative image relevancy scores from InstructBLIP and LLaVA given plain (a-c) black, (d-f) white, and (g-i) noise images with the query “Please describe the image.” The distributions of relevancy scores are nearly identical regardless of the reference input image, supporting the robustness of the VTC module.



(a)

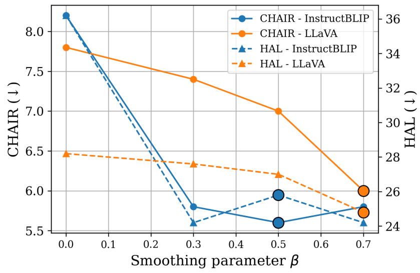


(b)

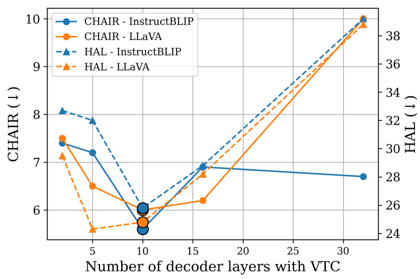
Figure 6: Ablation study results for InstructBLIP on the (a) CHAIR and (b) AMBER benchmarks. The plots show the performance of the baseline, VTC-only, AAR-only, and full CAAC (VTC + AAR) settings in terms of hallucination rates and recall metrics.

or truncated sequences, likely due to significant changes in attention distribution causing information loss in later layers. We thus examined the effect of varying the number of layers, from the first 2 to all 32 decoder layers. The best performance, with minimal hallucination rates, was observed when VTC was applied to the first 10 layers, as shown in Figure 7b.

We also assessed the smoothing parameter β, testing values from 0 to 0.7. A value of $\beta = 1$ mirrored the issues seen with all-layers VTC application and was excluded. Intermediate values (0.3 to 0.7) reduced hallucination rates most effectively, with $\beta = 0.5$ and $\beta = 0.7$ yielding the best results for InstructBLIP and LLaVA, respectively (Figure 7a). These observations suggest that modest early-to-mid-layer calibration can encourage



(a)



(b)

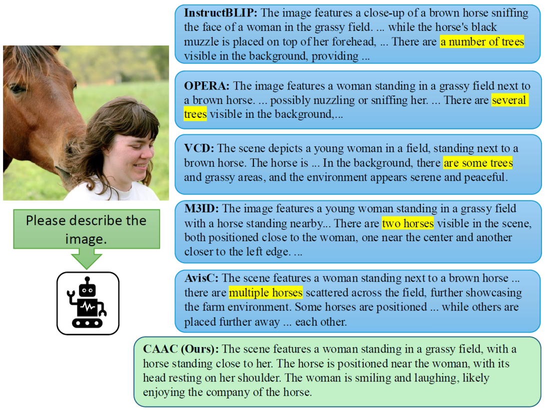
Figure 7: Distribution of image-token relevancy scores at the decoder’s last layer given a black canvas and the query "Please describe the image.". A pronounced skew toward a few image tokens for (a) LLaVA and (b) InstructBLIP can be witnessed.

a balanced attention to the image while allowing the later layers to promote semantically relevant tokens.

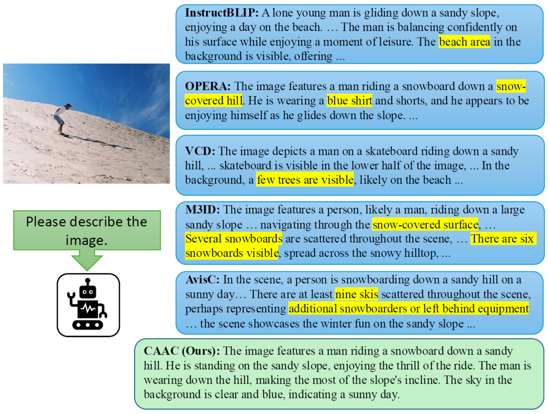
E Qualitative Evaluation

Additional examples of CAAC's performance compared to baselines (e.g., LLaVA, OPERA) on the AMBER dataset are presented in Figure **??**. These cases further illustrate CAAC's ability to reduce hallucinations across varied image-instruction pairs, with hallucinated content marked in red.

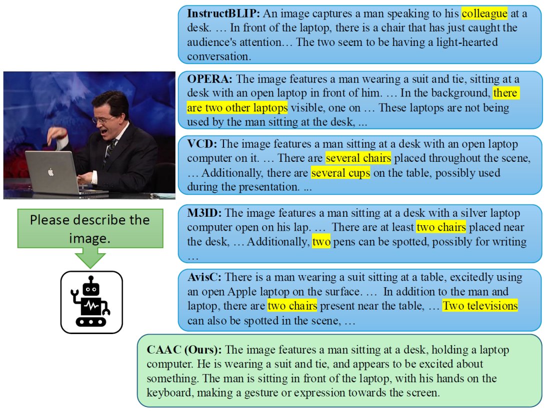
article [utf8]inputenc graphicx subcaption



(a) InstructBLIP: Case 1

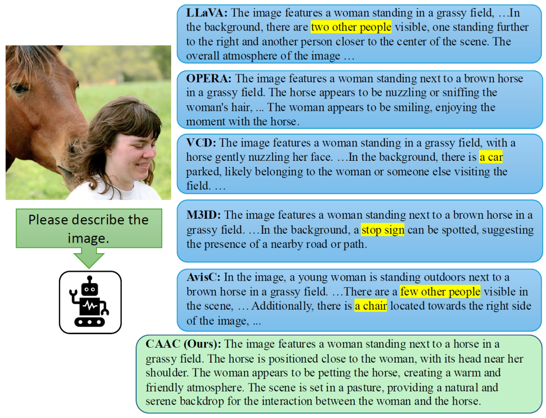


(b) InstructBLIP: Case 2

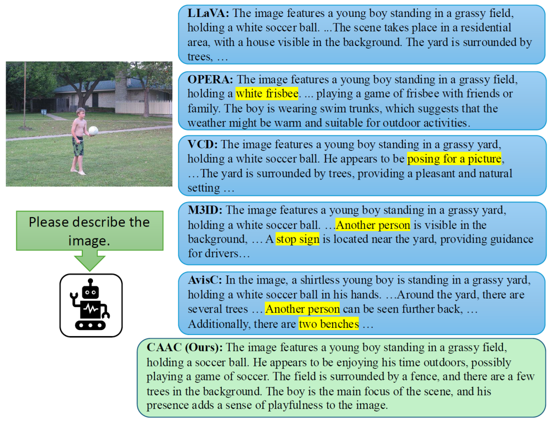


(c) InstructBLIP: Case 3

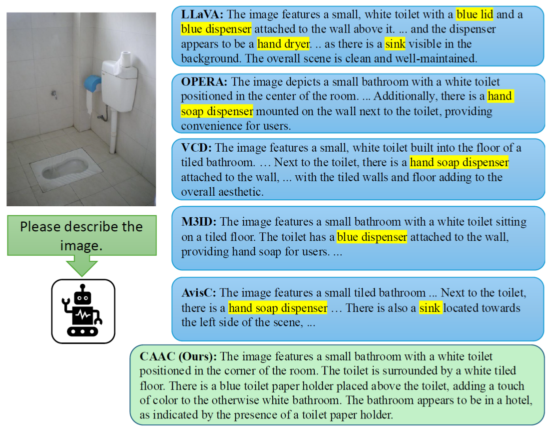
Figure 8: Comparison of CAAC outputs with baseline methods for the InstructBLIP model. Hallucinations are highlighted.



(a) LLaVA: Case 1



(b) LLaVA: Case 2



(c) LLaVA: Case 3

Figure 9: Comparison of CAAC outputs with baseline methods for the LLaVA model. Hallucinations are highlighted.