See Only When Needed: Context-Aware Attention Intervention for Mitigating Hallucinations in LVLMs

Yuqing Lei[1], Wenbo Lyu[1], Yingjun Du[2], Xiantong Zhen[3], Cees G.M. Snoek[2], and Ling Shao[1]*

[1] University of Chinese Academy of Sciences
[2] University of Amsterdam
[3] United Imaging Healthcare Co., Ltd.

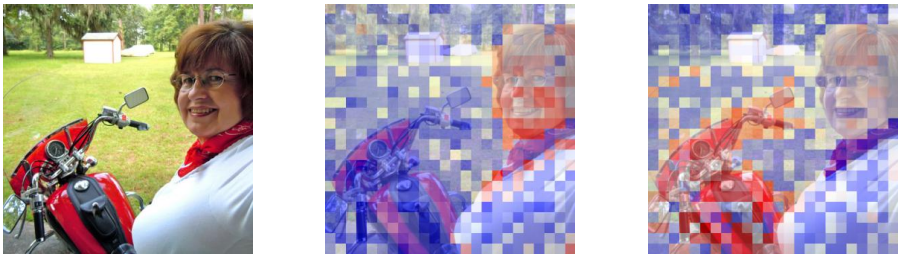
Abstract. Large Vision-Language Models (LVLMs) excel at multimodal tasks but remain prone to object hallucinations. Prior training-free remedies often *uniformly* strengthen visual signals, which may also amplify irrelevant regions and introduce spurious evidence, harming fluency. We propose **Context-aware Attention Intervention (CAI)**, a training-free inference-time mechanism that enforces a *see only when needed* principle via *two-axis selectivity*: *where* to look and *when* to intervene. At each decoding step, CAI derives token-specific visual relevance from early-layer representations to localize semantically aligned regions, and applies a conservative, entropy- and depth-gated attention tilt only for uncertainty-spiking tokens in deeper layers where visual grounding degrades, leaving confident tokens and irrelevant regions largely unchanged. This targeted intervention strengthens visual grounding while preserving linguistic fluency, and it yields consistent improvements *even without* contrastive decoding, which remains optional as an auxiliary bias-suppression module. Extensive experiments across multiple LVLM backbones and benchmarks show that CAI achieves state-of-the-art hallucination mitigation, and our analysis characterizes CAI as a KL-minimal attention reweighting with bounded interference under inactive gates or small tilts. Code is available at `https://github.com/Iris1946/CAI`.

Keywords: Vision-Language Models · Hallucination Mitigation · Attention Intervention

1 Introduction

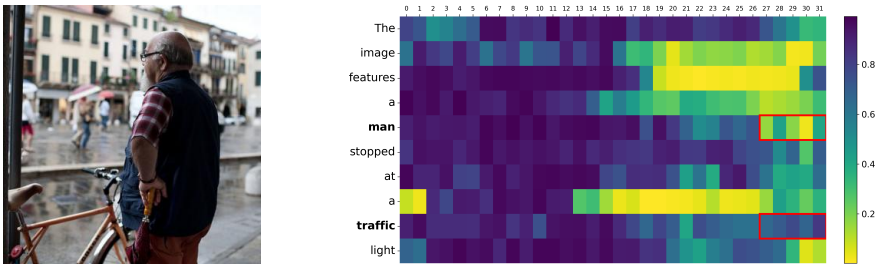
Large Vision-Language Models (LVLMs) [2, 10, 26, 43, 52] have achieved remarkable performance on multimodal tasks such as image captioning [19], visual question answering [10,26], and multimodal reasoning [13,29,34,37,50]. Despite these advances, LVLMs remain prone to hallucinations [3, 21, 24, 51], producing outputs that are linguistically plausible yet ungrounded in the visual input. This

* Corresponding author: ling.shao@ieee.org



(a) (b) (c)

Fig. 1: Visualization of token-image similarity. Regions highlighted in red indicate higher relevance between generated tokens and visual content. Given visual input (a) and the query *"Please describe the image in detail"*, region (b) is most associated when generating *"woman"*, whereas region (c) is most relevant when generating *"motorcycle"*.



(a) (b)

Fig. 2: Given visual input (a) and the query *"Please describe the image in detail"*, (b) shows **the evolution of token entropy across decoding layers**. In deeper layers, hallucination-prone tokens (e.g., *"traffic light"*) exhibit markedly higher entropy than grounded tokens (e.g., *"man"*), whereas tokens dominated by language priors (e.g., *"The", "a", "at"*) remain low-entropy.

failure mode critically undermines reliability in real-world applications where correctness and visual grounding are essential.

Prior studies [21,32,51] attribute hallucinations to statistical biases in large-scale training data, including frequent objects and co-occurrence patterns. Hallucinations also arise from model-intrinsic factors, particularly the over-reliance on language priors inherited from a pretrained language model [12,17,18,32,42]. During autoregressive decoding, early mismatches between text and vision can be amplified, leading the model to commit to visually unsupported content.

Existing strategies for mitigating hallucinations broadly fall into training-based and training-free approaches. Training-based methods rely on curated datasets [23,45,47] for fine-tuning [7,16,47] or reinforcement learning [36,46,49], aiming to correct biases via additional supervision. In contrast, training-free

methods [1, 6, 11, 18, 27, 28, 41, 53] operate purely at inference time to re-balance vision and language signals, offering a practical remedy without retraining cost.

A common design in training-free methods is to *uniformly* boost visual signals, e.g., increasing attention to image tokens across decoding steps and tokens. While effective in some cases, uniform amplification can also elevate irrelevant regions and spurious correlations, introducing new errors and degrading fluency. This motivates a *non-interference* requirement for inference-time intervention: the model does not need to "see more" everywhere; it should *see only when needed*, strengthening visual grounding only for tokens at risk while leaving confident tokens and irrelevant regions essentially unchanged.

Our analysis suggests that such selective intervention should follow two regularities. *First*, visual relevance is *token-specific*: different words should attend to different regions (Fig. 1). In the same image and prompt, "woman" and "motorcycle" correspond to distinct regions, indicating that strengthening vision must be spatially targeted rather than global. *Second*, hallucination risk is *depth- and uncertainty-dependent*: in deeper decoding layers, hallucination-prone tokens (e.g., "traffic light") exhibit substantially higher predictive entropy than grounded content tokens (e.g., "man"), whereas function words dominated by language priors (e.g., "The", "a", "at") remain low-entropy (Fig. 2). This makes token entropy a natural *risk proxy* for triggering intervention and, equally importantly, a mechanism to explicitly control *when not to intervene*. Together, these observations indicate that an effective remedy must be selective along two axes: (i) *where* to look—localize regions by token-specific relevance; and (ii) *when* to intervene—activate intervention only under high uncertainty and in deeper layers where visual grounding tends to degrade.

Based on these principles, we propose **C**ontext-aware **A**ttention **I**ntervention (**CAI**), a *training-free* mechanism that performs *minimal* and *token-conditioned* attention intervention at inference time to enforce "see only when needed". At each decoding step, CAI derives token-specific visual relevance from early-layer representations to locate semantically aligned regions, and then *conditionally* applies an attention tilt toward these regions only for uncertainty-spiking tokens in deeper layers, leaving low-entropy tokens and shallow layers largely untouched. As an optional add-on, we also report results with contrastive decoding following PAI [28] to further suppress text-only biases; importantly, CAI itself yields consistent improvements without relying on decoding tricks. Empirical evaluations on LLaVA-1.5 [25], InstructBLIP [10], and Qwen-VL [2] show that CAI consistently reduces hallucinations and improves performance across benchmarks.

Beyond empirical gains, we provide a principled characterization of CAI to address *why* and *when* such selective intervention is beneficial. Specifically, CAI can be viewed as a KL-minimal attention reweighting that increases token–image relevance with bounded perturbation; for high-uncertainty tokens, small tilts reduce the negative log-likelihood under an alignment condition between the induced visual shift and the local descent direction; and depth/entropy gating yields bounded (often negligible) interference when the gate is inactive.

Contributions. (1) We introduce CAI, a training-free *two-axis selective* attention intervention that decides *where* to look (token-specific visual relevance) and *when* to intervene (uncertainty- and depth-gated), enforcing a non-interference "see only when needed" principle. (2) We present an early-to-late grounding transfer mechanism: early-layer relevance localizes aligned regions, while intervention is applied only in deeper layers where visual grounding degrades, avoiding uniform amplification and preserving fluency. (3) We provide a principled theoretical characterization of CAI as a KL-minimal attention reweighting, together with conditions under which small tilts improve likelihood for high-uncertainty tokens and guarantees of bounded interference when the gate is inactive.

2 Related Work

Large Vision-Language Models (LVLMs). In recent LVLMs [2,10,26,43,52], vision-language integration enables Large Language Models (LLMs) [2,5,8,31,40] to extend their reasoning beyond text by incorporating visual information. Images are first processed by a vision encoder into embeddings that are aligned with the LLM's textual space [19, 20, 35], allowing the model to interpret visual content [19], answer questions [10, 26], and generate multimodal outputs [13, 29, 34, 37, 50]. This integration effectively empowers LLMs to perform tasks that require understanding both language and vision, bridging the gap between seeing and reasoning. Despite their capabilities, LVLMs often suffer from object hallucination [3, 21, 24, 51], describing objects that are not present in the image. Such errors reduce reliability, particularly in tasks requiring precise visual understanding. Rather than changing architectures or retraining on curated data, we act *at inference time*. Our approach selectively reinforces attention to token-relevant visual evidence, improving fidelity while remaining training-free and plug-and-play across LVLM backbones.

Mitigating hallucinations in LVLMs. Recent LVLM research has highlighted object hallucination as a persistent challenge. Hallucinations arise from both statistical biases in large-scale training data, such as frequently occurring objects and common object co-occurrences [21, 32, 51], and model-intrinsic factors, notably the reliance on language priors from pretrained language models [12, 17, 18, 32, 42]. LVLM outputs often align with language priors despite contradicting visual evidence. Strategies to mitigate hallucinations generally fall into training-based and training-free approaches. Training-based methods employ curated datasets [23, 45, 47] for fine-tuning [7, 16, 47] or reinforcement learning [36, 46, 49] to reduce bias-induced errors, but their high computational cost limits scalability. In contrast, training-free methods [1, 6, 11, 18, 27, 28, 41, 53] intervene at inference time, enhancing visual grounding to counteract the model's over-reliance on language priors, offering a more efficient alternative for improving output fidelity. Inspired by Neo et al. [30], who show that visual information is localized near object tokens, we propose a *token-level, depth- and uncertainty-gated* method: it identifies token-image relevance and activates only when predictive entropy spikes in deeper layers. This design strengthens grounding precisely

where needed and complements contrastive decoding to counter language-prior bias.

3 Preliminary

LVLMs generalize Large Language Models (LLMs) to enable joint reasoning over textual and visual modalities. A vision encoder extracts visual features $\mathbf{v} = [v_1, \ldots, v_{N_v}]$ from an input image, while a language model encodes textual input into query tokens $\mathbf{x} = [x_1, \ldots, x_{N_x}]$. These modalities are integrated through mechanisms such as multilayer perceptrons [25] or Q-Formers [10], generating a compact representation that conditions the language model. This fused representation facilitates autoregressive generation, formalized as:

$$y_t \sim p\left(y_t|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}\right) = \mathcal{S}\left(f_\theta\left(y_t|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}\right)\right). \tag{1}$$

where y_t denotes the token generated at step t, $\mathbf{y}_{<t}$ represents the preceding token sequence, and $\mathcal{S}$ is the softmax operator over the vocabulary. Here, f_θ corresponds to the language model parameterized by θ. The model is implemented as a stack of transformer blocks, each comprising multi-head self-attention (MHA) and a feed-forward network (FFN). The attention operation for head n is:

$$\mathrm{Attn}_n(h) = \mathcal{S}\left(\mathbf{A}_n\right) V_n, \quad \mathbf{A}_n = \frac{Q_n K_n^\top}{\sqrt{d_k}}. \tag{2}$$

where $Q_n, K_n, V_n \in \mathbb{R}^{N \times d_k}$ are the query, key, value projections of the hidden state h, and d_k is the key dimensionality, $N = N_x + N_v$ represents the total number of multimodal tokens. The attention weights $\mathbf{A}_n \in \mathbb{R}^{N \times N}$ capture token-to-token dependencies, facilitating contextual feature mixing. The outputs of all H heads are concatenated and projected through an output matrix W_o:

$$\mathrm{MHA}(h) = \mathrm{Concat}\left(\mathrm{Attn}_1\left(h\right), \cdots, \mathrm{Attn}_H\left(h\right)\right) \cdot W_o. \tag{3}$$

Finally, each transformer block applies an FFN to the MHA output, introducing nonlinear transformations that enhance contextual embeddings.

4 Method

In this section, we present **Context-aware Attention Intervention (CAI)**, a training-free approach to mitigate hallucinations at inference time. Previous training-free methods [1, 6, 11, 18, 27, 28, 41] typically intervene *uniformly* across visual tokens, which may introduce noise and spurious correlations from irrelevant regions. In contrast, CAI enforces a *"see only when needed"* behavior by combining three components: (i) a token-image similarity module that locates semantically relevant regions for each decoding token; (ii) an entropy- and depth-gated attention intervention that selectively amplifies attention to those regions only when hallucination risk is high; and (iii) an optional contrastive decoding step that further suppresses language-only continuations. The overall pipeline of CAI is illustrated in Fig. 3, and summarized in Algorithm 1.

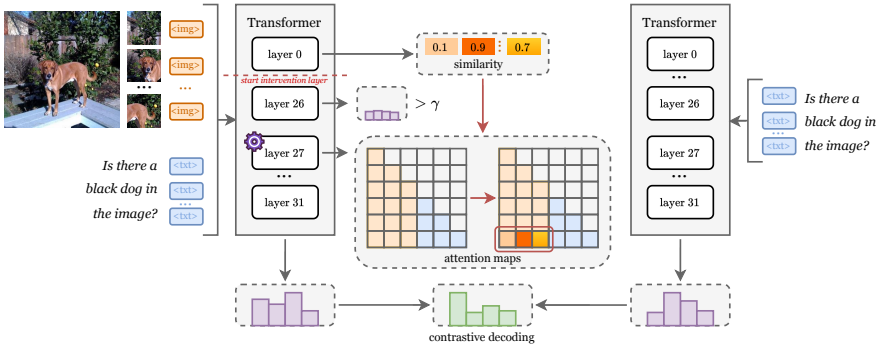


Fig. 3: Overview of CAI. At each decoding step, the lowest layer evaluates the similarity between the generated token and each patch token (Eq. 4), yielding a visual grounding signal. Attention interventions are applied to deep decoding layers with high entropy (Eq. 6), redirecting attention toward patch tokens in proportion to their similarity (Eq. 5). Integrated with contrastive decoding (Eq. 7), CAI mitigates reliance on language priors and reinforces visual grounding to reduce hallucinations.

4.1 Token-Aware Visual Grounding

Token-image similarity. At step t, CAI measures the similarity between the current text token and the visual patch tokens. Specifically, we extract the hidden state of the last token, $h_t^0 \in \mathbb{R}^{d_h}$, from the 0-th layer, where representations remain relatively disentangled from higher-order semantic abstractions (as shown in Section 6). We then compute the inner product between this $L2$-normalized hidden state and the N_v visual patch features, $\mathbf{v} \in \mathbb{R}^{N_v \times d_h}$, to derive the grounding weights:

$$\mathbf{w}_t^i = \sigma\left(\hat{\mathbf{v}}_i^\top \cdot \hat{h}_t^0\right), \quad \text{where} \quad \hat{h}_t^0 = \frac{h_t^0}{\|h_t^0\|_2}, \ \hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2}. \tag{4}$$

Here, $\sigma(\cdot)$ denotes the sigmoid function, which constrains the similarity scores to $(0, 1)$, yielding $\mathbf{w}_t \in \mathbb{R}^{N_v}$. These weights capture fine-grained text-vision alignment and provide a stable grounding signal to guide attention in deeper layers, enhancing multimodal consistency.

Similarity-guided attention intervention. Based on the similarity $\mathbf{w}_t$, CAI performs an attention intervention on the image regions indexed by $[i_s, i_e)$ during the generation of the t-th token:

$$\mathbf{A}_{t,i_s:i_e} = \mathbf{A}_{t,i_s:i_e} + |\mathbf{A}_{t,i_s:i_e}| \odot \mathbf{w}_t, \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{H \times N \times N}$ denotes the multi-head attention. The slice $\mathbf{A}_{t,i_s:i_e} \in \mathbb{R}^{H \times N_v}$, with $N_v = i_e - i_s$, represents the attention weights from the t-th token to N_v image tokens across H heads. The operator $\odot$ indicates element-wise multiplication. By scaling the attention magnitude $|\mathbf{A}_{t,i_s:i_e}|$ with the similarity $\mathbf{w}_t$,

Algorithm 1 Context-aware Attention Intervention (CAI)

Input: Transformer layers L, query embedding $\mathbf{x}$, image feature $\mathbf{v}$, image start token i_s, image end token i_e, start intervention layer l_s, entropy threshold γ, decoding coefficient λ.

Output: Response token y_t at decoding step t.

for $l \in [0, L)$ **do**
  if $l = 0$ **then**
    $\mathbf{w}_t \leftarrow \sigma(\mathbf{v} \cdot h_t^l)$. % token-image similarity
  end if
  % conditional attention intervention
  if $l \geq l_s$ **and** $-\sum p(h_t^{l-1}) \log p(h_t^{l-1}) > \gamma$ **then**
    $\mathbf{A}^l_{t,i_s:i_e} \leftarrow \mathbf{A}^l_{t,i_s:i_e} + \left|\mathbf{A}^l_{t,i_s:i_e}\right| \odot \mathbf{w}_t$.
  end if
  $\bar{h}_t^l \leftarrow h_t^l + \mathrm{MHA}_l(h_t^l)$.
  $h_t^{l+1} \leftarrow h_t^l + \mathrm{FFN}_l(\bar{h}_t^l)$.
end for
$p(y_t|\mathbf{v}, \mathbf{x}) \leftarrow \lambda\, p(y_t|\mathbf{v}, \mathbf{x}) - (\lambda - 1)\, p(y_t|\mathbf{x})$. % contrastive decoding

the intervention amplifies attention proportionally to semantic relevance. This mechanism reinforces context-aware visual grounding by directing the model's focus toward the image regions that are most pertinent to the token under generation.

4.2 Entropy- and Depth-Gated Intervention

The indiscriminate application of interventions across all tokens and layers risks degrading multimodal representations that are already adequately grounded. To mitigate this, CAI employs a dual-gating mechanism:

$$-\sum p(h_t^l) \log p(h_t^l) > \gamma, \quad \text{subject to} \quad l > l_s. \tag{6}$$

First, interventions are strictly confined to layers deeper than l_s, where visual signals typically degrades. Second, CAI selectively targets tokens with high hallucination risk by evaluating their predictive uncertainty. Specifically, we project the layer-normalized h_t^l through the language modeling head (LMHead) to compute the predictive distribution: $p(h_t^l) = \mathrm{Softmax}\left(\mathrm{LMHead}\left(\mathrm{LayerNorm}\left(h_t^l\right)\right)\right)$. Interventions are triggered only if the entropy of this distribution exceeds γ. By conditioning on both layer depth and entropy, CAI focuses reinforcement on hallucination-prone tokens, ensuring that interventions are precise and effective. We further analyze this entropy- and depth-gated design in Section 5, showing that CAI implements a KL-minimal attention tilt and yields provable gains for high-entropy tokens under small tilts.

4.3 Contrastive Decoding

In CAI, contrastive decoding is introduced as an optional component. Following PAI [28], hallucinations from over-reliance on language priors are mitigated by

contrasting multimodal prediction $p(y_t|\mathbf{v}, \mathbf{x})$ with text-only prediction $p(y_t|\mathbf{x})$:

$$\hat{p}(y_t|\mathbf{v}, \mathbf{x}) = \lambda\, p(y_t|\mathbf{v}, \mathbf{x}) - (\lambda - 1)\, p(y_t|\mathbf{x}), \tag{7}$$

where λ is a contrastive decoding coefficient. This penalizes text-only hypotheses to encourage visual grounding. In contrast to PAI [28], where attention intervention alone yields limited impact and contrastive decoding drives the significant gains reported in Section 6, our CAI achieves substantial hallucination mitigation with the standalone attention intervention mechanism (CAI†, i.e., $\lambda = 1$), making contrastive decoding optional.

5 Theoretical Analysis

We next provide a formal analysis of CAI by viewing it as an *exponential tilt* of the baseline attention. Under this lens, CAI enjoys four desirable properties that mirror our design principles. First, among all ways of modifying attention within a fixed KL budget, the CAI update is the *KL-minimal* reweighting that most increases expected token–image similarity (Theorem 1). Second, for high-entropy tokens, small tilts that shift visual evidence in the NLL descent direction provably reduce the loss (Theorem 2), matching our entropy-gated intervention. Third, under a simple model where visual information decays with depth, later layers yield larger marginal benefits from reinforcement (Theorem 3), justifying depth-gated intervention. Finally, for small tilts or weak similarity, the change in attention is tightly bounded in total variation (Theorem 4), ensuring that CAI behaves as a low-risk, "see only when needed" modification.

Setup. Let $a \in \Delta^{N-1}$ be the attention over N visual tokens at the current step and $s \in \mathbb{R}^N_{\geq 0}$ the token-image similarity. When gated by $\mathbb{I}[H_t^{(l-1)} > \gamma] \cdot \mathbb{I}[\, l \geq l_0\,]$, CAI applies a multiplicative tilt:

$$\tilde{a}_i \;=\; \frac{a_i\, \exp(\lambda s_i)}{\sum_j a_j\, \exp(\lambda s_j)}, \qquad \lambda \geq 0,$$

else $\tilde{a} = a$. Let $x(a) = \sum_i a_i v_i$ be the aggregated visual evidence and logits $z_y = u_y + w_y^\top x(a)$, with NLL $\mathcal{L}(a) = -\log \mathrm{softmax}(z)_{y^\star}$.

Theorem 1 (KL-minimality of CAI tilting). *For any baseline a and similarity s, the distribution $q^\star(\lambda) \propto a \odot \exp(\lambda s)$ uniquely solves $\max_{q \in \Delta^{N-1}} \mathbb{E}_q[s]$ s.t. $D_{\mathrm{KL}}(q \,\|\, a) \leq \varepsilon$ for some $\lambda \geq 0$ meeting the KL budget. Thus, CAI realizes the least-change reweighting that raises expected relevance.*

Theorem 2 (Entropy-gated improvement). *Assume the linearized logit model $z_y = u_y + w_y^\top x(a)$ and local smoothness of $\log \sum_y e^{z_y}$. Let $H = -\sum_y p_y \log p_y$ with $p = \mathrm{softmax}(z)$. If $H \geq H_0 > 0$ and the CAI direction aligns with the NLL descent, i.e., $\langle g(a), \Delta x \rangle > 0$ where $g(a) = \sum_y (p_y - \mathbb{1}[y{=}y^\star]) w_y$ and $\Delta x = x(\tilde{a}) - x(a)$, then there exists $\lambda_0 > 0$ such that for all $0 < \lambda \leq \lambda_0$, $\mathcal{L}(\tilde{a}) < \mathcal{L}(a)$. Hence high-entropy tokens are the regime where CAI is guaranteed to help for sufficiently small tilts.*

Theorem 3 (Depth advantage via visual decay). *Suppose the visual component of hidden states evolves as* $x^{(l)} = M^{(l)}x^{(l-1)}$ *with* $\mathbb{E}\,\rho(M^{(l)}) \le \rho < 1$. *Then for* $l \ge l_0$, $\|x^{(l)}\| \le \rho^{l-l_0}\|x^{(l_0)}\|$. *Therefore the signal-to-noise ratio of visual evidence decays geometrically with depth, making depth-gated intervention* ($l \ge l_0$) *yield larger marginal returns.*

Theorem 4 (Non-interference under small tilts). *For small* λ, $D_{\mathrm{KL}}(\tilde{a}\,\|\,a) = \frac{1}{2}\lambda^2\mathrm{Var}_a[s] + o(\lambda^2)$; *by Pinsker,* $\|\tilde{a}-a\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\tilde{a}\,\|\,a)}$. *Thus, when the gate is off or* s *is weak, CAI perturbs attention only negligibly.*

Proofs of Theorems 1–4 are provided in Appendix.

6 Experiments

6.1 Experiment Setup

Datasets. Our evaluation employs three benchmarks. **POPE** [21] detects hallucinations via binary object-existence queries on MS-COCO [22], A-OKVQA [33], and GQA [14], using *random*, *popular*, and *adversarial* sampling to assess accuracy, memorization bias, and robustness. **CHAIR** [32] evaluates hallucinations in free-form captioning by quantifying both the proportion of hallucinated object instances and the proportion of captions containing hallucinations. **MME** [44] delivers a comprehensive evaluation across fourteen subtasks formulated as yes-or-no queries, encompassing perceptual dimensions such as object existence, count, position, and color.

Implementation details. Our evaluation is conducted on LLaVA-1.5 [25], InstructBLIP [10], and Qwen-VL [2]. Baselines, including VCD [18], PAI [28] and ONLY [41], are employed with their default configurations to ensure a fair comparison. For our approach, we perform a grid search over the hyperparameters l_s, γ and λ, while σ is instantiated as the sigmoid function. All experiments are executed on a single 80GB NVIDIA A800 GPU.

Table 1: CHAIR evaluation results for LLaVA-1.5 and InstructBLIP.

	Method	Max Token = 64		Max Token = 128	
		$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$
LLaVA-1.5	Regular	26.0	8.8	56.6	16.8
	VCD	23.8	8.2	59.6	16.6
	PAI	26.0	8.7	54.0	15.2
	ONLY	21.0	7.3	48.4	14.7
	CAI	**17.8**	**6.9**	**39.6**	**12.4**
InstructBLIP	Regular	29.0	10.2	54.2	16.7
	VCD	26.4	8.7	55.8	**15.9**
	PAI	24.6	8.3	56.8	16.7
	ONLY	25.2	8.8	55.8	17.4
	CAI	**24.2**	**8.2**	**52.8**	16.5

Table 2: POPE evaluation results for LLaVA-1.5, InstructBLIP, and Qwen-VL.

Dataset		Method	LLaVA-1.5		InstructBLIP		Qwen-VL	
			Acc ↑	*F1* ↑	*Acc* ↑	*F1* ↑	*Acc* ↑	*F1* ↑
MS-COCO	Random	Regular	85.46	86.07	82.82	83.56	85.09	83.46
		VCD	85.74	86.70	85.53	85.32	89.59	89.18
		PAI	86.67	87.26	85.43	83.60	85.74	84.20
		ONLY	88.38	88.41	87.32	87.77	88.97	88.37
		CAI	**89.24**	**89.14**	**89.55**	**89.35**	**89.90**	**89.49**
	Popular	Regular	81.20	82.22	75.77	77.69	84.13	81.97
		VCD	81.93	83.36	80.97	81.15	87.57	87.02
		PAI	82.77	83.60	77.13	79.14	85.10	83.09
		ONLY	85.00	85.23	77.87	79.93	87.23	86.31
		CAI	**86.03**	**85.53**	**84.30**	**83.65**	**88.30**	**87.70**
	Adversarial	Regular	75.87	78.27	74.23	76.61	81.50	79.51
		VCD	76.90	79.62	79.17	79.99	84.20	84.07
		PAI	76.83	79.14	74.87	77.53	83.33	81.64
		ONLY	79.93	81.18	75.57	78.29	84.23	83.67
		CAI	**82.77**	**82.73**	**81.83**	**81.56**	**84.97**	**84.69**
A-OKVQA	Random	Regular	82.07	83.31	81.53	82.86	86.53	85.19
		VCD	82.17	84.14	85.27	85.72	89.77	89.50
		PAI	83.33	84.85	83.10	84.46	87.03	85.80
		ONLY	86.03	86.93	85.67	86.70	89.23	88.78
		CAI	**89.00**	**89.03**	**89.87**	**89.73**	**90.07**	**89.91**
	Popular	Regular	75.30	78.99	74.80	77.98	86.00	84.78
		VCD	77.20	80.54	78.97	80.76	89.53	89.34
		PAI	76.37	79.79	76.47	79.61	87.37	86.17
		ONLY	79.40	81.86	76.73	80.07	88.83	88.46
		CAI	**84.60**	**85.23**	**84.87**	**85.40**	**89.53**	**89.43**
	Adversarial	Regular	67.07	73.70	68.33	73.89	81.03	80.36
		VCD	68.30	74.86	73.33	77.00	82.13	82.92
		PAI	67.60	74.23	68.67	74.55	82.10	81.58
		ONLY	70.47	75.86	68.77	74.95	82.20	82.74
		CAI	**77.40**	**79.79**	**75.23**	**78.04**	**82.60**	**83.20**
GQA	Random	Regular	82.03	83.86	80.17	81.55	83.83	82.60
		VCD	81.70	83.99	83.37	83.78	88.33	88.26
		PAI	83.37	85.03	81.23	82.67	85.90	84.77
		ONLY	86.20	87.20	83.87	85.02	88.13	87.54
		CAI	**89.10**	**89.15**	**88.10**	**87.84**	**90.13**	**89.87**
	Popular	Regular	71.93	76.88	72.27	75.97	80.77	80.15
		VCD	74.37	78.97	77.57	79.40	84.33	84.82
		PAI	72.73	77.60	73.77	77.34	82.67	81.89
		ONLY	75.93	79.62	74.80	78.41	82.80	82.72
		CAI	**83.43**	**84.39**	**81.20**	**82.05**	**84.97**	**84.95**
	Adversarial	Regular	67.93	74.35	68.33	73.25	79.20	78.96
		VCD	68.63	75.41	73.40	76.38	81.70	82.74
		PAI	69.00	75.34	69.10	74.19	80.87	80.41
		ONLY	71.50	76.75	69.37	74.92	81.43	81.67
		CAI	**78.33**	**80.46**	**76.00**	**77.65**	**83.27**	**83.51**

6.2 Results

Overall performance across benchmarks. Across POPE, CHAIR, and MME, CAI consistently improves hallucination mitigation and visual grounding over the strongest training-free baseline, and the gains hold across three representative LVLM backbones (LLaVA-1.5, InstructBLIP, and Qwen-VL), indicating that CAI is not model-specific but transfers reliably.

Hallucination mitigation on POPE and CHAIR. Table 2 shows consistent improvements on POPE across all backbones, reflecting fewer visually ungrounded responses under a controlled hallucination protocol. Table 1 further corroborates this trend on free-form captions: CAI reduces both sentence-level and instance-level hallucination rates, and the reduction remains under different maximum output lengths, suggesting that CAI mitigates error accumulation during autoregressive decoding.

General visual reasoning on MME. Beyond hallucination-specific metrics, Table 3 demonstrates improved performance on MME for both object-level (existence, count) and attribute-level (position, color) reasoning. This aligns with CAI's design—token-specific region localization strengthens fine-grained grounding, while entropy- and depth-gated intervention avoids uniform amplification of irrelevant regions, leading to more reliable object and attribute predictions.

Table 3: MME evaluation results for LLaVA-1.5, InstructBLIP, and Qwen-VL.

	Method	Object-level		Attribute-level		*Score* ↑
		Existence ↑	*Count* ↑	*Position* ↑	*Color* ↑	
LLaVA-1.5	Regular	185.00	126.67	128.33	148.33	588.33
	VCD	180.00	141.67	128.33	153.33	603.33
	PAI	185.00	131.67	133.33	153.33	603.33
	ONLY	190.00	141.67	133.33	163.33	628.33
	CAI	**190.00**	**143.33**	**148.33**	**178.33**	**660.00**
InstructBLIP	Regular	170.00	75.00	68.33	140.00	453.33
	VCD	155.00	78.33	76.67	155.00	465.00
	PAI	150.00	88.33	**78.33**	150.00	466.67
	ONLY	170.00	68.33	63.33	145.00	446.66
	CAI	**175.00**	**100.00**	70.00	**165.00**	**510.00**
Qwen-VL	Regular	165.00	135.00	163.33	175.00	638.33
	VCD	170.00	120.00	133.33	175.00	598.33
	PAI	175.00	135.00	163.33	175.00	648.33
	ONLY	180.00	136.67	153.33	180.00	650.00
	CAI	**180.00**	**146.67**	**163.33**	**185.00**	**675.00**

Table 4: Efficiency–performance trade-off on LLaVA-1.5. CAI† and PAI† uses attention intervention only, CAI and PAI additionally applies contrastive decoding.

Method	Latency ↓ (ms/token)	Throughput ↑ (token/ms)	GPU Memory ↓ (MB)	*POPE* ↑	*CHAIR* ↓	*MME* ↑
Regular	89.62	9.41	14241	74.81	56.6	588.33
VCD	215.22 (×2.40)	2.22 (×0.24)	15299 (×1.07)	75.89 (+1.08)	59.6 (+3.0)	603.33 (+15.00)
ONLY	116.51 (×1.30)	6.31 (×0.67)	14281 (×1.00)	78.63 (+3.82)	48.4 (−8.2)	628.33 (+40.00)
PAI†	90.97 (×1.02)	9.27 (×0.99)	14241 (×1.00)	74.57 (−0.24)	54.8 (−1.8)	581.66 (−6.67)
CAI†	**92.35** (×1.03)	**9.15** (×0.97)	**14251** (×1.00)	77.13 (+2.32)	43.6 (−13.0)	608.33 (+20.00)
PAI	123.57 (×1.38)	7.09 (×0.75)	14281 (×1.00)	75.77 (+0.96)	54.0 (−2.6)	603.33 (+15.00)
CAI	131.33 (×1.47)	6.8 (×0.72)	14291 (×1.00)	**83.67** (+8.86)	**39.6** (−17.0)	**660.00** (+71.67)

6.3 Discussion

Analysis of contrastive decoding. Table 4 summarizes the accuracy–efficiency trade-offs. The attention-only variant (CAI†) consistently outperforms the Regular baseline on POPE, CHAIR, and MME while maintaining comparable latency and throughput. This demonstrates that the principal gains stem from the proposed token- and risk-aware attention intervention, rather than from decoding tricks. In contrast, PAI†, which applies uniform amplification, yields limited or even negative improvements, supporting our hypothesis that indiscriminate boosting can introduce irrelevant evidence and impair grounding.

Adding contrastive decoding (CAI†→CAI) further improves performance across benchmarks with only marginal overhead. For fairness, CAI adopts the same contrastive decoding scheme as PAI; however, the results indicate that CAI does not depend on it for its core improvements. Instead, contrastive decoding serves as a complementary module that provides incremental precision, particularly under elevated hallucination risk. Overall, the key distinction lies in where, when, and how attention is modulated—through token-specific relevance estimation, entropy- and depth-based gating, and KL-minimal reweighting—while decoding operates as a shared auxiliary component rather than the primary source of gain.

Effect of deep-layer entropy in detecting hallucination. We adopt predictive entropy as a training-free, model-internal signal available at inference time and empirically validate its utility as a hallucination risk indicator. Predictive entropy is used only as a conservative trigger to avoid over-correction: high-entropy tokens concentrate hallucination risk in deeper layers (Fig. 4), so gating preserves fluency by leaving confident tokens untouched. Fig. 5 highlights a known failure mode: *confident hallucination*. In this case, entropy stays below the threshold, and CAI is not triggered. This limitation is inherent to uncertainty-based gat-

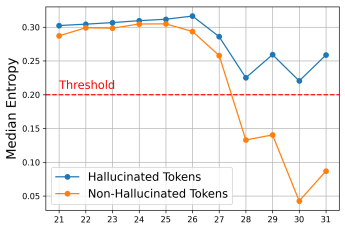
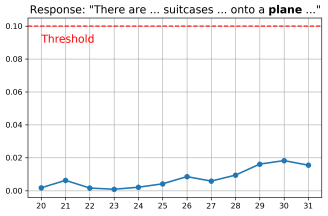


Fig. 4: Median entropy of tokens across layers. With an intervention threshold, hallucination and non-hallucination tokens are distinctly separable in the deeper layers.

(a) Image sample



(b) Layer entropy of "plane"

Fig. 5: Case study: failure mode of confident hallucination. Given the prompt "Please describe this image in detail", the model generates the hallucinated token "plane", which is not present in (a). Note that its predictive entropy across multiple layers remains below threshold, illustrating a failure mode where confident hallucinations evade entropy-based intervention.

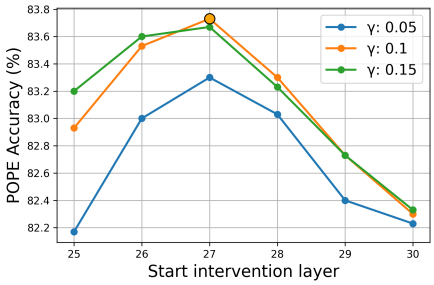


Fig. 6: Ablation study on the starting intervention layer l_s and the entropy threshold γ under the random setting of A-OKVQA in POPE.

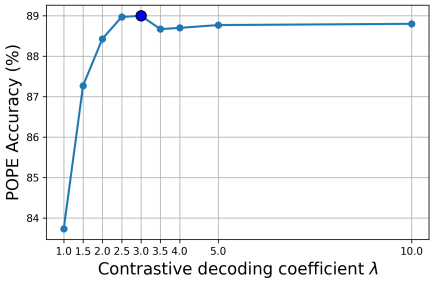
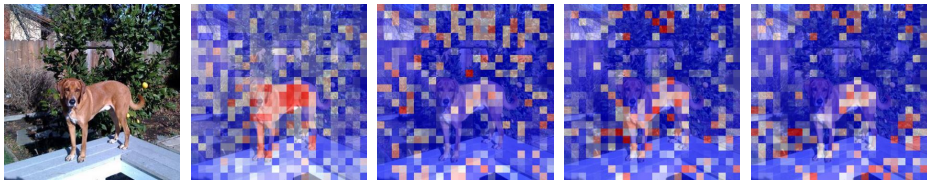


Fig. 7: Ablation study on the contrastive decoding coefficient λ under the random setting of A-OKVQA in POPE.



(a) Image sample **(b)** 0-th layer **(c)** 11-th layer **(d)** 21-th layer **(e)** 31-th layer

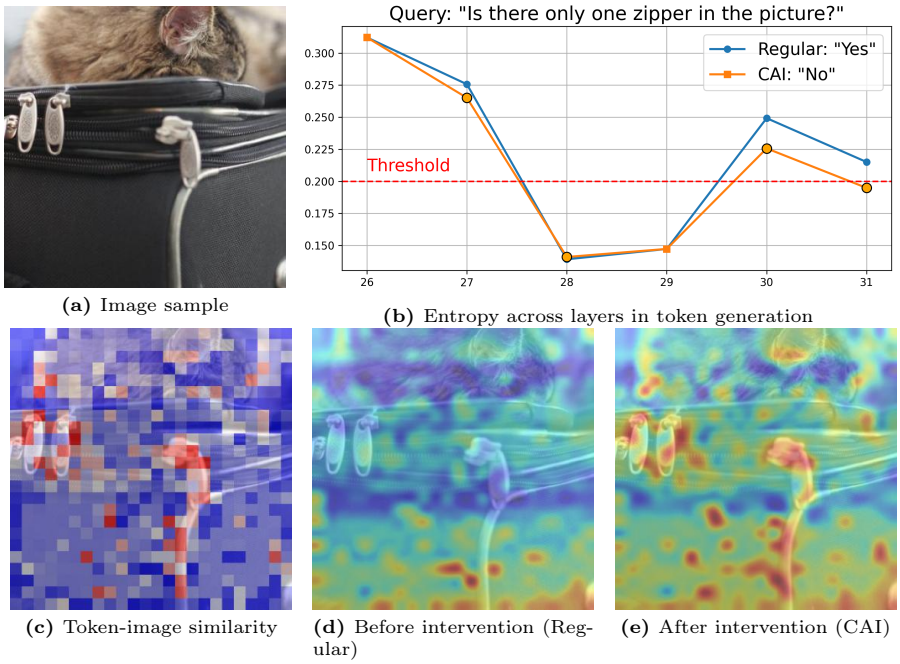
Fig. 8: Visualization of token-image similarity across decoding layers. Given visual input (a) and the query *"Is there a black dog in the image?"*. It can be observed that the hidden states in the 0-th layer strongly correspond to *'dog'*, whereas the deeper layers exhibit a diminished focus on local object features, reflecting a shift toward more global representations.

ing, and it motivates using contrastive decoding as an optional safeguard against text-only hypotheses.

Effect of l_s and γ in intervention conditions. We sweep the start layer $l_s \in [25, 30]$ and entropy threshold $\gamma \in \{0.05, 0.1, 0.15\}$ on A-OKVQA (random setting). Fig. 6 peaks at $l_s = 27$ and $\gamma = 0.1$. Intervening too *early* amplifies shallow-layer noise, while intervening too *late* misses error accumulation; similarly, a threshold that is too *low* over-triggers on easy tokens, and too *high* under-triggers on genuinely uncertain tokens. These trends support our design: decide *where* to look by token-image similarity, and *when* to act by depth and predictive entropy.

Effect of λ in contrastive decoding. We grid-search the decoding coefficient λ (Fig. 7); accuracy is maximized at $\lambda = 3.0$ on A-OKVQA (random setting). Small λ under-penalizes language-prior continuations; excessively large λ over-restricts decoding and can hurt fluency. We use $\lambda \approx 3$ when precision is prioritized, and $\lambda = 1$ (i.e., no contrastive decoding, $CAI^{\dagger}$) in strict latency budgets.

Effect of the lowest-layer representation. As shown in Fig. 8, the 0-th layer hidden states exhibit a strong spatial alignment with the target object



(a) Image sample

(b) Entropy across layers in token generation

(c) Token-image similarity

(d) Before intervention (Regular)

(e) After intervention (CAI)

Fig. 9: Case study of visual input (a) with the query *"Is there only one zipper in the picture?"*. The intervention layers are shown in (b), and token-image similarity in (c) establishes grounding between the response token and the visual input. The attention maps from the 0-th head in layer 31, before and after the similarity-guided intervention, are visualized in (d) and (e).

("dog"), successfully preserving fine-grained, localized visual features; in contrast, deeper layers shift focus toward more global and abstract representations, reducing sensitivity to individual object features. This observation aligns with recent interpretability studies on decoder-style LVLMs (e.g., LLaVA), which demonstrate that image-text communication [4] and object-localized information [30] are strongest in early layers, while middle-to-deep layers increasingly aggregate visual evidence into lexical abstractions [4] and become more influenced by language priors [30]. This structural progression naturally motivates our design: use early layers to estimate where the token-relevant visual evidence is, and intervene only in deeper layers where grounding tends to degrade.

Case study. Fig. 9 presents a case study of MME with a visual input (a) and the query *"Is there only one zipper in the picture?"*. (b) depicts the entropy of the hallucinated response token ("Yes") generated by LLaVA-1.5, contrasted with the non-hallucinated response token ("No") produced by our CAI . We set the start intervention layer $l_s = 26$ and the entropy threshold $\gamma = 0.2$, observing that the prediction entropy at layers $26, 27, 29,$ and 30 exceeds γ, indicating a heightened risk of hallucination. Accordingly, intervention is applied at the sub-

sequent layers, namely $27, 28, 30,$ and 31. During the intervention, token-image similarity (c) between the response token and input image (a) is computed at the lowest layer to establish a grounding baseline. In the high-risk layers, the attention maps visualized in (d) often lack attention to the relevant regions. By amplifying the attention weights of vision regions guided by (c), the attention map in (e) is reoriented toward token-relevant image regions and thereby mitigates hallucination.

7 Conclusion

We presented Context-aware Attention Intervention (CAI), a training-free and plug-and-play approach for mitigating hallucinations in large vision-language models via a *non-interference* principle: *see only when needed*. Rather than uniformly amplifying visual signals, CAI performs a *minimal* and *token-conditioned* attention intervention that is activated only for high-risk tokens and primarily in deeper layers where visual grounding tends to degrade, thereby strengthening grounding while preserving fluency. We further provide a principled characterization of CAI as a KL-minimal attention reweighting, together with conditions under which small tilts improve likelihood for high-uncertainty tokens and guarantees of bounded interference when the gate is inactive. Extensive results across multiple LVLM backbones and benchmarks demonstrate consistent hallucination reduction with lightweight inference-time overhead.

References

1. An, W., Tian, F., Leng, S., Nie, J., Lin, H., Wang, Q., Chen, P., Zhang, X., Lu, S.: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29915–29926 (2025)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
4. Basu, S., Grayson, M., Morrison, C., Nushi, B., Feizi, S., Massiceti, D.: Understanding information storage and transfer in multi-modal large language models. Advances in Neural Information Processing Systems **37**, 7400–7426 (2024)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Chen, J., Zhang, T., Huang, S., Niu, Y., Zhang, L., Wen, L., Hu, X.: Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4209–4221 (2025)
7. Chen, Z., Zhu, Y., Zhan, Y., Li, Z., Zhao, C., Wang, J., Tang, M.: Mitigating hallucination in visual language models with visual supervision. arXiv preprint arXiv:2311.16479 (2023)

8. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
9. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J.R., He, P.: Dola: Decoding by contrasting layers improves factuality in large language models. In: International Conference on Learning Representations. vol. 2024, pp. 54158–54183 (2024)
10. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
11. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14303–14312 (2024)
12. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14375–14385 (2024)
13. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
14. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)
15. Huo, F., Xu, W., Zhang, Z., Wang, H., Chen, Z., Zhao, P.: Self-introspective decoding: Alleviating hallucinations for large vision-language models. arXiv preprint arXiv:2408.02032 (2024)
16. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27036–27046 (2024)
17. Lee, S., Park, S.H., Jo, Y., Seo, M.: Volcano: mitigating multimodal hallucination through self-feedback guided revision. Proceedings of the 2023 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (2023)
18. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: 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. pp. 13872–13882 (2024)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
20. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
21. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023)

22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
23. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
24. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
27. Liu, S., Ye, H., Zou, J.: Reducing hallucinations in large vision-language models via latent space steering. In: The Thirteenth International Conference on Learning Representations (2025)
28. Liu, S., Zheng, K., Chen, W.: Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In: European Conference on Computer Vision. pp. 125–140. Springer (2024)
29. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025)
30. Neo, C., Ong, L., Torr, P., Geva, M., Krueger, D., Barez, F.: Towards interpreting visual information processing in vision-language models. In: The Thirteenth International Conference on Learning Representations (2025)
31. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
32. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing p. 4035–4045 (2018)
33. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022)
34. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
35. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7464–7473 (2019)
36. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. arXiv preprint arXiv:2309.14525 (2023)
37. Tan, H., Ji, Y., Hao, X., Lin, M., Wang, P., Wang, Z., Zhang, S.: Reason-rft: Reinforcement fine-tuning for visual reasoning. arXiv preprint arXiv:2503.20752 (2025)
38. Tang, F., Liu, C., Xu, Z., Hu, M., Huang, Z., Xue, H., Chen, Z., Peng, Z., Yang, Z., Zhou, S., et al.: Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26147–26159 (2025)

39. Team, Q.: Qwen3.5: Accelerating productivity with native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5
40. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
41. Wan, Z., Zhang, C., Yong, S., Ma, M.Q., Stepputtis, S., Morency, L.P., Ramanan, D., Sycara, K., Xie, Y.: Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. Proceedings of the International Conference on Computer Vision (2025)
42. Wu, Y., Zhao, Y., Zhao, S., Zhang, Y., Yuan, X., Zhao, G., Jiang, N.: Overcoming language priors in visual question answering via distinguishing superficially similar instances. Proceedings of the 29th International Conference on Computational Linguistics p. 5721–5729 (2022)
43. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
44. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
45. Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., Zhuang, Y.: Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12944–12953 (2024)
46. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024)
47. Yue, Z., Zhang, L., Jin, Q.: Less is more: Mitigating multimodal hallucination from an eos decision perspective. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics pp. 11766–11781 (2024)
48. Zhang, X., Zhu, Y., Gu, C., Yuan, X., Zhao, Q., Cao, J., Tang, F., Fan, S., Shen, Y., Shen, C., et al.: Hallucination begins where saliency drops. arXiv preprint arXiv:2601.20279 (2026)
49. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839 (2023)
50. Zhou, H., Li, X., Wang, R., Cheng, M., Zhou, T., Hsieh, C.J.: R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. arXiv preprint arXiv:2503.05132 (2025)
51. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. The Eleventh International Conference on Learning Representations (2023)
52. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
53. Zou, X., Wang, Y., Yan, Y., Lyu, Y., Zheng, K., Huang, S., Chen, J., Jiang, P., Liu, J., Tang, C., Hu, X.: Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. The Forty-second International Conference on Machine Learning (ICML) (2025)

Appendix

A Proofs for Theoretical Results

(Proof of Theorem 1). Consider $\max_{q\in\Delta^{N-1}} \mathbb{E}_q[s]$ s.t. $D_{\mathrm{KL}}(q\|a) \le \varepsilon$. The Lagrangian with multiplier $\eta \ge 0$ for the KL constraint and τ for normalization is

$$\mathcal{L}(q,\eta,\tau) = \sum_i q_i s_i - \eta\left(\sum_i q_i \log\frac{q_i}{a_i} - \varepsilon\right) + \tau\left(\sum_i q_i - 1\right).$$

First-order optimality (KKT) gives, for each i, $s_i - \eta\,(\log q_i - \log a_i + 1) + \tau = 0$ so that $\log q_i = \log a_i + \frac{1}{\eta}s_i - \frac{\tau+\eta}{\eta}$, hence $q_i \propto a_i \exp(\frac{1}{\eta}s_i)$. Writing $\lambda = \frac{1}{\eta} \ge 0$ yields the exponential tilt $q^\star(\lambda) \propto a \odot e^{\lambda s}$. Since $D_{\mathrm{KL}}(\cdot\|a)$ is strictly convex in q and the objective is linear, the solution is unique. Moreover, with $A(\lambda) = \log\sum_i a_i e^{\lambda s_i}$, one has $D_{\mathrm{KL}}(q^\star(\lambda)\|a) = \lambda A'(\lambda) - A(\lambda)$ and $\frac{d}{d\lambda} D_{\mathrm{KL}}(q^\star(\lambda)\|a) = \lambda A''(\lambda) = \lambda \operatorname{Var}_{q^\star(\lambda)}[s] \ge 0$, so the KL budget ε selects a unique $\lambda \ge 0$ by monotonicity.

(Proof of Theorem 2). Let $x(a) = \sum_i a_i v_i$ and $z_y = u_y + w_y^\top x(a)$, $p = \mathrm{softmax}(z)$, and $\mathcal{L}(a) = -\log p_{y^\star}$. The gradient of $\mathcal{L}$ w.r.t. x is $\nabla_x \mathcal{L}(a) = \sum_y p_y w_y - w_{y^\star} =: g(a)$. By the descent lemma for L-smooth functions (log-sum-exp is smooth), for $\Delta x = x(\tilde{a}) - x(a)$,

$$\mathcal{L}(\tilde{a}) - \mathcal{L}(a) \;\le\; \langle g(a), \Delta x\rangle + \frac{L}{2}\|\Delta x\|^2,$$

for some local $L > 0$ depending on $\{w_y\}$ and p. If the alignment condition is met (equivalently, $\langle -g(a), \Delta x\rangle > 0$, i.e., $\langle g(a), \Delta x\rangle < 0$), then the linear term is negative. Because the CAI tilt depends smoothly on λ and $\Delta x = O(\lambda)$ for $\lambda \to 0$, there exists $\lambda_0 > 0$ such that the (negative) linear term dominates the quadratic remainder for all $0 < \lambda \le \lambda_0$, implying $\mathcal{L}(\tilde{a}) < \mathcal{L}(a)$. The entropy condition $H \ge H_0 > 0$ guarantees we are in the uncertain regime where $g(a) \ne 0$ (non-vanishing gradient), so the alignment condition is meaningful.

(Proof of Theorem 3). By assumption, the visual component evolves as $x^{(l)} = M^{(l)} x^{(l-1)}$ and there exists $\rho < 1$ such that, for all $l \ge l_0 + 1$, the mixing is contractive.[4] Applying sub-multiplicativity of operator norms,

$$\begin{aligned}\|x^{(l)}\| &= \left\|M^{(l)} M^{(l-1)} \cdots M^{(l_0+1)}\, x^{(l_0)}\right\| \\ &\le \prod_{k=l_0+1}^{l} \|M^{(k)}\|\, \|x^{(l_0)}\| \\ &\le \rho^{\,l-l_0}\, \|x^{(l_0)}\|.\end{aligned}$$

Hence the visual signal decays geometrically with depth, making deeper layers yield larger marginal gains from reinforcement.

[4] If we further assume the spectral (operator) norm and that $M^{(l)}$ are normal, the bound $\|M^{(l)}\| \le \rho(M^{(l)}) \le \rho$ follows from $\rho(M) \le \|M\|$. In general, the same conclusion holds under the standard bounded-mixing assumption $\|M^{(l)}\| \le \rho < 1$.

(Proof of Theorem 4). For the tilted distribution $q^\star(\lambda)$ with log-normalizer $A(\lambda) = \log \sum_i a_i e^{\lambda s_i}$,

$$D_{\mathrm{KL}}(q^\star(\lambda)\|a) = \mathbb{E}_{q^\star(\lambda)}\left[\log \frac{q^\star(\lambda)}{a}\right] = \lambda A'(\lambda) - A(\lambda).$$

A Taylor expansion of A at $\lambda = 0$ gives $A(\lambda) = \mu\lambda + \frac{1}{2}\sigma^2\lambda^2 + O(\lambda^3)$ with $\mu = \mathbb{E}_a[s]$ and $\sigma^2 = \mathrm{Var}_a[s]$. Since $A'(\lambda) = \mu + \sigma^2\lambda + O(\lambda^2)$,

$$\begin{aligned} D_{\mathrm{KL}}(q^\star(\lambda)\|a) &= \lambda(\mu + \sigma^2\lambda + O(\lambda^2)) - (\mu\lambda + \tfrac{1}{2}\sigma^2\lambda^2 + O(\lambda^3)) \\ &= \tfrac{1}{2}\sigma^2\lambda^2 + O(\lambda^3). \end{aligned}$$

By Pinsker's inequality, $\|q^\star(\lambda) - a\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(q^\star(\lambda)\|a)} = O(\lambda)$, and any Lipschitz functional of a (e.g., $w_y^\top x(a)$) is perturbed only by $O(\lambda)$. Therefore, when the gate is off or s is weak (yielding small λ), the effect of CAI is negligible.

B Additional Experiments

B.1 Evaluation metrics

CHAIR evaluates hallucination in image captioning by quantifying references to objects absent from the image. Using 500 randomly sampled MS-COCO images and the prompt *"Please describe this image in detail"*, we compute two complementary metrics. Sentence-level CHAIR denotes the proportion of captions containing at least one hallucinated object:

$$\mathrm{CHAIR}_S = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}, \tag{8}$$

and instance-level CHAIR is the proportion of hallucinated mentions among all object mentions:

$$\mathrm{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}. \tag{9}$$

POPE serves as a binary classification task. For each image, the model is prompted with *"Is [object] in this image?"* to determine object presence. Performance is measured using standard metrics: *accuracy*, the proportion of correct predictions, and *F1-score*, the harmonic mean of precision and recall:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{10}$$

MME quantifies fidelity via yes-or-no questions on existence, count, position, and color. For each object, the model's responses are compared to ground truth, and the overall *MME-score* aggregates correctness across all attributes, providing a fine-grained measure of object-level hallucination and semantic alignment.

B.2 Additional Results

Comparison with more baselines. We extended our comparative analysis to include recent methods evaluated on LLaVA-1.5 using the POPE benchmark. As demonstrated in Table 5, CAI achieves the highest average accuracy and F1 scores. Importantly, this superior performance is coupled with substantial computational efficiency. The next best-performing baseline, LVLMs-Saliency, necessitates both saliency model inference and per-token attention reallocation. Consequently, it incurs a latency of 4294.36 ms/token, rendering it 33.7× slower than CAI (131.33 ms/token).

Table 5: Additional POPE evaluation results for LLaVA-1.5.

Method	Random		Popular		Adversarial		*Average*		Latency ↑
	Acc↑	*F1*↑	*Acc*↑	*F1*↑	*Acc*↑	*F1*↑	*Acc*↑	*F1*↑	(ms/token)
LLaVA-1.5	85.46	86.07	81.20	82.22	75.87	78.27	80.84	82.19	89.62
DoLa [9]	81.77	82.09	89.17	88.54	**86.37**	**85.98**	85.77	85.54	-
SID [15]	80.33	81.38	**89.47**	**89.10**	85.87	85.88	85.22	85.45	-
FarSight [38]	89.07	89.01	85.63	86.04	79.30	81.05	84.67	85.37	-
LVLMs-Saliency [48]	89.10	88.20	87.07	86.01	83.33	83.12	86.50	85.77	4294.36
CAI	**89.24**	**89.14**	86.03	85.53	82.77	82.73	**86.68**	**85.80**	131.33

Evaluation on a stronger backbone. To validate generalization to newer architectures, we evaluate CAI on Qwen3.5-4B [39], which employs a hybrid design. Specifically, the model utilizes Softmax Attention at layers 3, 7, 11, 15, 19, 23, 27, and 31, while applying Gated Linear Attention (GLA) to all remaining layers. Since GLA operates on recurrent states rather than explicit token-to-token attention maps, CAI's similarity-guided intervention has no direct attention map to modulate. We therefore intervene only on deep Softmax-Attention layers. As reported in Table 6, directly intervening on the GLA gates (denoted as $\text{CAI}_{\text{GLA-gated}}$) yields inconsistent gains. In contrast, applying CAI exclusively to the Softmax layers robustly improves performance on both the CHAIR and MME

Table 6: Additional CHAIR and MME evaluation results for Qwen-3.5.

Method	Max Token = 64		Max Token = 128		*MME* ↑
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	
Qwen-3.5	18.6	9.6	36.2	12.6	685.00
PAI	18.4	9.1	37.2	12.3	691.66
$\textbf{CAI}_{\text{GLA-gated}}$	17.2	10.3	35.0	14.4	700.00
CAI	**16.6**	**8.5**	**29.8**	**11.4**	**720.00**

benchmarks, establishing a new baseline across all metrics. Notably, these evaluations utilize a single, fixed hyperparameter configuration across both datasets, without benchmark-specific re-tuning.

B.3 Additional Discussions

Implementation details of contrastive decoding. Following PAI [28], $p(y_t|\mathbf{x})$ in Eq. 7 is obtained by a *text-only* forward pass (i.e., removing the image token; for LLaVA: text around "<*image*>", for InstructBLIP/Qwen-VL: "*{question} Answer:*"). This adds a small overhead of 0.04 s/token.

Efficiency of sparse token interventions. As shown in Fig. 10, in LLaVA on the MME benchmark with $l_s = 26$, CAI intervenes on only $6.67\% \sim 39.38\%$ of tokens in layers 27 to 31, with interventions declining in deeper layers. This empirically validates the bounded-risk property (Theorem 4), confirming that the method perturbs only a small subset of tokens, thereby reducing hallucinations while maintaining high inference efficiency.

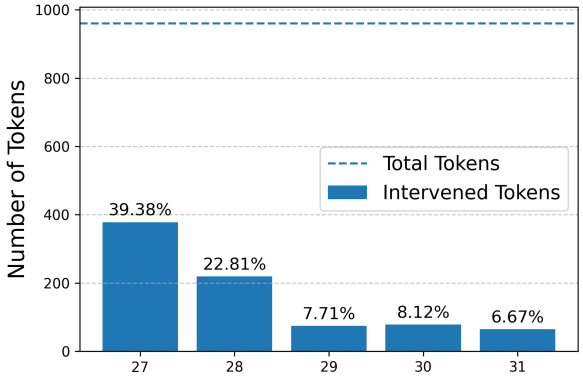


Fig. 10: Number of tokens requiring intervention at each layer, along with their relative proportion to the total number of generated tokens.

Hyperparameter sensitivity and adaptive rules. CAI primarily relies on grid-search tuning for the start intervention layer l_s, the entropy threshold γ, and the contrastive decoding coefficient λ. However, fixed hyperparameters can directly transfer to newer architectures (Table 6) without benchmark-specific re-tuning. To eliminate grid search entirely, we also add label-free adaptive defaults to reduce grid search: use a depth-proportional start layer $l_s = \lfloor 0.85L \rfloor$ or, for hybrid architectures, the deepest Softmax-Attention layers; and use an adaptive entropy threshold $\gamma = \mu_H^{(t)} + 0.5\sigma_H^{(t)}$, where $\mu_H^{(t)}$ and $\sigma_H^{(t)}$ are running entropy

statistics from previously decoded tokens in the same sample. This requires no validation labels. We will also clarify that contrastive decoding is optional: the attention-only CAI† already improves over Regular and PAI† in Table 4, while $\lambda > 1$ only adds complementary language-prior suppression.

Additional case study. Fig. 11 illustrates three cases of long-text generation. The CHAIR metric is used to identify hallucinated and ground-truth tokens in "Regular" (LLaVA-1.5) responses. By applying CAI to reinforce visual grounding, the model amplifies the attention of response tokens toward relevant visual tokens, thereby mitigating hallucinations. Note that, due to space constraints, only the attention map of a specific head in the intervention layer is visualized here.

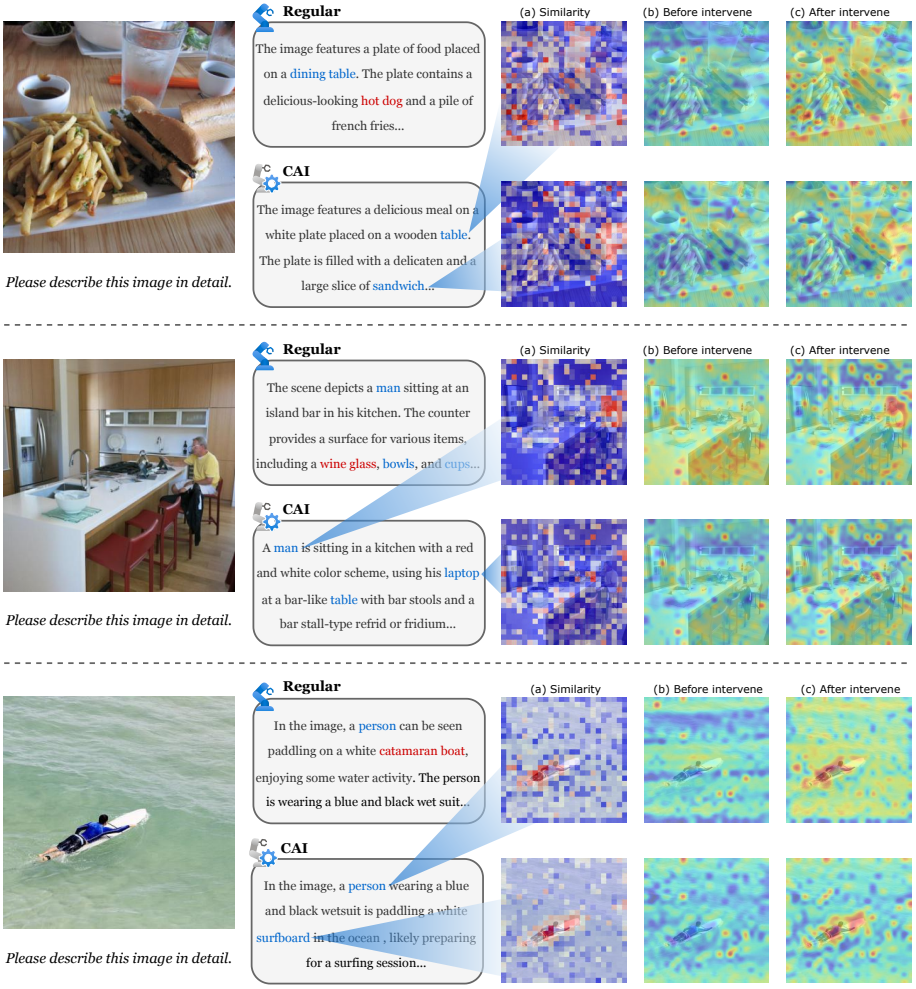


Fig. 11: Case study of CHAIR in long text generation. Given the image and the prompt on the left, "Regular" response corresponds to LLaVA-1.5, with red denoting CHAIR hallucination words and blue representing correctly recognized terms. Employing "CAI" effectively mitigates hallucinations, with the attention intervention process illustrated in (a), (b), and (c) on the right.