Mitigating Multimodal Hallucinations via Gradient-based Self-Reflection

Shan Wang[1,2,3] Maying Shen[1] Nadine Chang[1] Chuong Nguyen[3] Hongdong Li[2]
Jose M. Alvarez[1]
[1]NVIDIA [2]Australian National University [3]Data61, CSIRO

Abstract

Multimodal large language models (MLLMs) achieve strong performance across diverse tasks but remain prone to hallucinations, where outputs are not grounded in visual inputs. This issue can be attributed to two main biases: text–visual bias, the overreliance on prompts and prior outputs, and co-occurrence bias, spurious correlations between frequently paired objects. We propose Gradient-based Influence-Aware Constrained Decoding (GACD), an inference-based method, that addresses both biases without auxiliary models, and is readily applicable to existing models without finetuning. The core of our approach is bias estimation, which uses first-order Taylor gradients to understand the contribution of individual tokens—visual features and text tokens—to the current output. Based on this analysis, GACD mitigates hallucinations through two components: (1) suppressing spurious visual features correlated with the output objects, and (2) rebalancing cross-modal contributions by strengthening visual features relative to text. Experiments across multiple benchmarks demonstrate that GACD effectively reduces hallucinations and improves the visual grounding of MLLMs outputs.

1. Introduction

Recent advances in Multimodal Large Language Models (MLLMs) show strong ability to produce coherent and context-aware content across a wide range of domains [2, 6, 9, 29, 46]. Despite their impressive advancements, these models remain prone to hallucination, wherein the generated text is not faithfully grounded in the visual modality [27, 37]. This limitation poses a critical barrier to establishing trust in the outputs of MLLMs.

The hallucinations observed in MLLMs can be largely attributed to two fundamental biases [20, 23, 27]. **Text-visual bias** refers to the excessive reliance on textual information—such as the input prompt and previously generated outputs—while neglecting the visual modality during generation. This bias becomes particularly pronounced in longer sequences, where MLLMs tend to depend more heavily on prior text and increasingly disregard visual cues [12, 54]. **Co-occurrence bias** arises from spurious statistical correlations embedded in the training data, which lead models to erroneously predict the presence of non-existent objects based on their frequent co-occurrence with observed objects in the visual inputs [27]. This bias is particularly challenging to mitigate, and existing approaches largely rely on statistical priors rather than offering statistically agnostic solutions [20, 55].

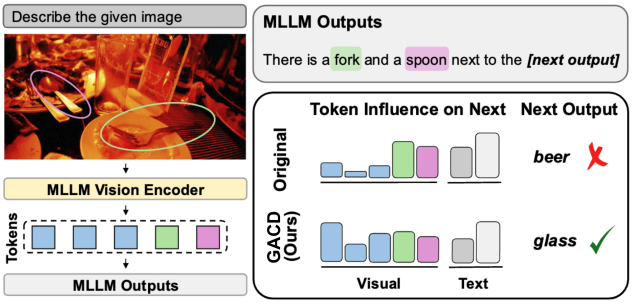


Figure 1. Overview of our influence-aware constrained decoding framework, which mitigates hallucinations by regulating token-level influence. It reduces text–visual bias by enhancing visual token influence (blue bars) in alignment with the most influential text inputs— prompts (gray) or previous outputs (white). It further mitigates co-occurrence bias through anchor-specific suppression, selectively suppressing visual tokens (green, magenta) anchored to previously emitted nouns.

Existing efforts to mitigate hallucinations in MLLMs can be broadly categorized into inference-based methods, which operate at the decoding stage [7, 12, 24, 34, 44], and training-based approaches, which intervene during model optimization [3, 5, 18, 20, 40, 48]. Inference-based approaches are valued for their cost-effectiveness, as they avoid the need for additional data collection, data bias examination, or extensive model retraining. However, these methods offer limited insight into the severity of underlying biases, leaving the root causes of hallucination insufficiently understood. In addition, some inference-based methods rely on auxiliary models—such as segmentation networks [7], detection systems [19] , or even additional MLLMs [10, 36, 45]— which introduce extra sources of er-

ror, depend on task-specific supervision.

Another limitation of existing methods lies in their lack of granularity when adjusting the underlying biases in MLLMs. Most approaches rely on heuristically tuned priors, which vary across datasets and fail to generalize reliably [24, 53]. Moreover, they apply uniform weighting across all visual features, offering no mechanism to selectively adjust bias at the level of individual features [32, 52]. This coarse treatment limits their effectiveness in mitigating co-occurrence bias, which arises from spurious statistical correlations between objects that are often represented by distinct visual features.

In this work, we propose an inference-based method that simultaneously addresses both text–visual bias and co-occurrence bias, without relying on auxiliary models or external supervision. The core of our approach is the estimation of underlying bias, achieved by quantifying the contribution of individual tokens—both visual features and text tokens—through gradients derived from a first-order Taylor expansion. Building on this analysis, the method mitigates hallucinations by reweighting tokens via two key components: (1) suppressing the influence of visual features that exhibit strong spurious correlations with the current output token, thereby reducing co-occurrence bias; and (2) rebalancing cross-modal contributions by enhancing the role of visual features to align more closely with that of text tokens in generating the current output. As illustrated in Fig. 1, our method, GACD, corrects hallucinated predictions—such as the spurious generation of "beer" in the presence of "fork" and "spoon"—by amplifying the contributions of visual tokens unrelated to those nouns, leading to outputs that are more faithfully grounded in the visual modality. Note also that our method is readily applicable to existing MLLMs at inference time.

We summarize our main contributions as follows.

- We introduce an inference-based method for hallucination mitigation in MLLMs, built on a principled estimation of underlying bias via gradients obtained from a first-order Taylor expansion. This estimation provides a mechanism for understanding and granularly adjusting their influences of individual visual features and text tokens on the generation of the current output token, all without requiring auxiliary models or finetuning.
- We design two complementary modules: (i) suppression of spurious visual features correlated with the current output token to alleviate co-occurrence bias, and (ii) cross-modal rebalancing to enhance the contributions of visual features relative to text tokens, thereby addressing text–visual bias.
- Extensive experiments demonstrate that GACD mitigates hallucinations and enhances accuracy while maintaining a favorable balance between accuracy and informativeness. GACD achieves up to 8% increase in overall score on AMBER [43], an 8% F1 boost on POPE [27], up to 45% improvement in detailness and a 92% accuracy gain on LLaVA-QA90 [30].

2. Related Work

Hallucination and Bias. Hallucinations in LLMs often arise from biases in the training data [16, 33], while in MLLMs, studies [13, 27, 42] show that hallucinations are closely linked to biases like text-visual and co-occurrence biases. Additionally, biases related to output position, which increase the risk of hallucination as output length grows, have been examined in [12, 54]. Existing methods [13, 23, 27] typically report only overall statistics, lacking a mathematical, sample-wise bias measurement. This distinction is important, as biases can vary case by case. Our approach measures sample-dependent bias via token-level gradient sensitivities, revealing how pre-trained MLLM parameters embed these biases [15, 22], and enabling self-reflective hallucination mitigation.

Hallucination Mitigation. Training-related hallucination mitigation methods [4, 18, 35, 48, 49] are expensive, requiring access to training data and specialized statistical analysis. Among them, LPOI [49] also employs an object-aware framework, highlighting the effectiveness of modeling object-level information for mitigating hallucinations. Reinforcement-learning approaches [10, 45, 50] rely on supplementary feedback, often from human annotators or auxiliary LLMs/MLLMs, and the latter may themselves hallucinate. By contrast, post-decoding techniques modify model logits at inference time without further training or external feedback, making them lightweight add-ons. In text-only LLMs, such methods aim to align outputs with factual knowledge [8, 25]. In MLLMs, post-decoding strategies emphasize the role of visual inputs [12, 24, 53] and can be classified into image-level and token-level interventions. Image-level decoding methods [32, 52] treat all objects in the input image uniformly, limiting their effectiveness in addressing co-occurrence hallucinations. Existing token-level methods either rely on external segmentation [7] and detection models [19] or lack awareness of object-related decoupling [44]. Moreover, these methods typically introduce an implicit trade-off between accuracy and informativeness, reducing hallucinations at the expense of omitting valid details. Attention-based methods [41, 51] require careful selection of specific layers and often introduce model-specific adjustments or heuristics. In contrast, our GACD directly estimates embedded bias and decouples object-aware visual tokens, enabling sample-specific hallucination mitigation without external data, models, or model-specific adjustments, while achieving a more favorable balance between accuracy and informativeness.

3. Method

In this section, we provide background on MLLMs, introduce the concept of token influence, and explain how GACD balances token influence to mitigate hallucinations.

3.1. Background on MLLMs

MLLMs generate a finite token sequence $\mathbf{y} = [y_1, \ldots, y_M]$ in response to a visual input (image or video) and a textual prompt. Let $\mathcal{V}$ be a finite vocabulary. The prompt is tokenized as $\mathbf{t}^p = [t_1^p, \ldots, t_N^p]$ with $t_n^p \in \mathcal{V}$. The visual input is encoded by a visual encoder into features, which are then projected into the token-embedding space $\mathbb{R}^d$, yielding visual tokens $\mathbf{t}^v = [t_1^v, \ldots, t_S^v]$ with $t_s^v \in \mathbb{R}^d$, where d is the shared token embedding dimension used for $\mathcal{V}$.

A MLLM with parameters θ computes, at each decoding step m, a logit vector

$$\mathbf{z}_m \;=\; \pi_\theta(\mathbf{t}^v,\, \mathbf{t}^p,\, \mathbf{y}_{<m}) \in \mathbb{R}^{|\mathcal{V}|}, \tag{1}$$

where $\mathbf{y}_{<m} = [y_1, \ldots, y_{m-1}]$ (empty when $m = 1$). This induces a categorical next-token distribution via the softmax $\sigma : \mathbb{R}^{|\mathcal{V}|} \to \Delta^{|\mathcal{V}|-1}$:

$$p_\theta(y_m \mid \mathbf{t}^v, \mathbf{t}^p, \mathbf{y}_{<m}) \;=\; \big[\sigma(\mathbf{z}_m)\big]_{y_m}, \quad 1 \le m \le M, \tag{2}$$

where $\sigma(\mathbf{z}_m) \in \Delta^{|\mathcal{V}|-1}$ denotes the probability distribution[1] over the vocabulary, and $[\cdot]_{y_m}$ selects the component corresponding to token $y_m \in \mathcal{V}$. At inference, y_m is sampled from this categorical distribution (e.g., greedy, beam search). The sequence likelihood factorizes by the chain rule:

$$p_\theta(\mathbf{y} \mid \mathbf{t}^v, \mathbf{t}^p) \;=\; \prod_{m=1}^{M} p_\theta(y_m \mid \mathbf{t}^v, \mathbf{t}^p, \mathbf{y}_{<m})\,. \tag{3}$$

Given a dataset $\mathcal{D}$ of $(\mathbf{t}^v, \mathbf{t}^p, \mathbf{y})$, maximum-likelihood training (or fine-tuning) estimates $\theta^\star$ by maximizing the conditional log-likelihood. Pretrained MLLMs encode statistical regularities (including spurious correlations) from training data in $\theta^\star$; such behavior can be probed without changing $\theta^\star$ via parameter-dependent analyses (e.g., gradients/attributions or counterfactual decodings) [15, 22], enabling self-reflective bias interpretation.

3.2. Gradient-Based Token Influence Estimation

To capture these embeded biases, we interpret how each input token perturbs the output logits. Let $\mathbf{z}_m^\star \in \mathbb{R}^{|\mathcal{V}|}$ denote the step-m logits $\mathbf{z}_m^\star = \pi_{\theta^\star}(\mathbf{t}^v, \mathbf{t}^p, \mathbf{y}_{<m})$. Around a reference sample point $(\mathbf{t}^{v(0)}, \mathbf{t}^{p(0)}, \mathbf{y}_{<m}^{(0)})$, the first-order Taylor expansion [39] of the logits $\mathbf{z}_m^\star$ is

$$\mathbf{z}_m^\star \;\approx\; \sum_{s=1}^{S} \mathbf{g}_{ms}^v t_s^v + \sum_{n=1}^{N} \mathbf{g}_{mn}^p t_n^p + \sum_{i=1}^{m-1} \mathbf{g}_{mi}^y y_i + Const, \tag{4}$$

where $Const$ denotes other terms that are constant w.r.t., $\mathbf{t}^v$ and $\mathbf{t}^p$ and the token-wise Jacobians are

$$\begin{aligned} \mathbf{g}_{ms}^v \;&\coloneqq\; \frac{\partial\, \mathbf{z}_m^\star}{\partial\, \mathbf{t}_s^v}\Big|_{\mathbf{t}^v = \mathbf{t}^{v(0)}}, \qquad \mathbf{g}_{mn}^p \;\coloneqq\; \frac{\partial\, \mathbf{z}_m^\star}{\partial\, \mathbf{t}_n^p}\Big|_{\mathbf{t}^p = \mathbf{t}^{p(0)}}, \\ \mathbf{g}_{mi}^y \;&\coloneqq\; \frac{\partial\, \mathbf{z}_m^\star}{\partial\, \mathbf{y}_i}\Big|_{\mathbf{y} = \mathbf{y}_{<m}^{(0)}}, \end{aligned} \tag{5}$$

where $\big|_{\cdot}$ indicate evaluation at the reference sample point. Taylor expansion details are in supplementary Sec. 1. Each $\mathbf{g}_{ms}^v$, $\mathbf{g}_{mn}^p$, $\mathbf{g}_{mi}^y$ indicate a small token perturbation in its embedding space to a perturbation of the predict logit vector in $\mathbb{R}^{|\mathcal{V}|}$. Following [38], we approximate the importance of each input token by the Manhattan norm of its gradient:

$$I_{ms}^v \;=\; \|\mathbf{g}_{ms}^v\|_1, \quad I_{mn}^p \;=\; \|\mathbf{g}_{mn}^p\|_1, \quad I_{mi}^y \;=\; \|\mathbf{g}_{mi}^y\|_1, \tag{6}$$

and $I_{ms}^v[c]$ represents the gradient from the output vocabulary c with respect to each visual tokens. Aggregating over tokens yields step-m group-level influences:

$$\mathtt{I}_m^v \;=\; \sum_{s=1}^{S} I_{ms}^v, \quad \mathtt{I}_m^p \;=\; \sum_{n=1}^{N} I_{mn}^p, \quad \mathtt{I}_m^y \;=\; \sum_{i=1}^{m-1} I_{mi}^y. \tag{7}$$

These quantities decompose, at the sample level, how visual tokens, prompt tokens, and prior outputs contribute to the logit of y_m, enabling interpretation of bias per sample.

3.3. Influence-Aware Constrained Decoding

GACD builds on token influence estimation with two components: (i) Object-aware Visual Token Grouping and (ii) Anchor-specific Influence-weighted Decoding. At step m, the former partitions visual tokens into object-related $\mathbf{t}^o$ and unrelated $\mathbf{t}^u$ based on objects detected in $\mathbf{y}_{<m}$. The latter extends contrastive decoding [26] by forming *Anchor-specific* negative guidance logits from pre-mentioned objects and computing a decoding weight α_m from token-influence measurements.

Object-aware Visual Token Grouping. For each step m, we detect nouns in $\mathbf{y}_{<m}$ and treat each noun y_i as an object mention. To link a mention to visual evidence, we measure the influence I_{is}^v of visual token s on step i, For every noun y_i, the visual token with maximal influence is selected to form a mask $\mathcal{M}_{is}$. The cumulative object mask at step m aggregates all prior noun-linked tokens:

$$\mathcal{M}_{ms} = \mathbf{1}\left[\sum_{i=1}^{m-1} \mathcal{M}_{is} > 0\right], \tag{8}$$

[1] We use "confidence" to denote the model-assigned probability of the emitted token.

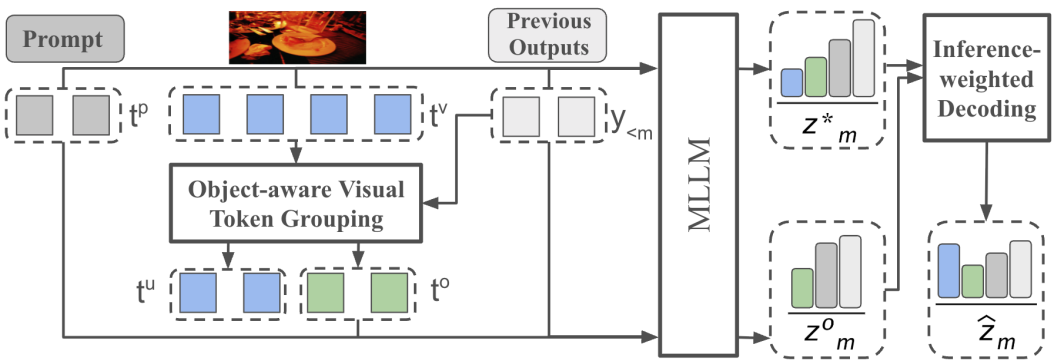


Figure 2. Overview of GACD. The method comprises (i) Object-aware Visual Token Grouping and (ii) Anchor-specific Influence-Weighted Decoding. At step m, previously mentioned objects are detected from $\mathbf{y}_{<m}$; visual tokens are partitioned into object-related textcolordarkgreen$\mathbf{t}^o$ and unrelated $\mathbf{t}^u$ via token influence (Sec. 3.2). Anchor-specific Influence-weighted Decoding extends contrastive decoding with token influence, explicitly amplifying the influence of $\mathbf{t}^u$ to jointly counter text-visual and co-occurrence biases; negative-guidance logits $\mathbf{z}^o_m$ are generated from $\{\mathbf{t}^o, \mathbf{t}^p, \mathbf{y}_{<m}\}$ to suppress text tokens and anchor-specific visual cues. Grouping is invoked only for noun prediction (where co-occurrence arises between object pairs); for non-noun prediction, we set $\mathbf{t}^o = \varnothing$ and uniformly amplify all visual tokens to balance text–visual bias.

where $\mathcal{M}_{is} = \mathbf{1}\big[\, y_i \in \text{Noun} \;\wedge\; s = \arg\max_j I^v_{ij} \,\big]$ and $\mathbf{1}[\cdot]$ is the indicator and $\wedge$ is logical AND.

The mask $\mathcal{M}_{ms}$ identifies visual tokens linked to nouns emitted before m. We then partition the visual tokens into *object-related* ($\mathbf{t}^o$) and *unrelated-to-objects* ($\mathbf{t}^u$) sets via a Hadamard product:

$$\mathbf{t}^o = \mathbf{t}^v \odot \mathcal{M}_m, \quad \mathbf{t}^u = \mathbf{t}^v \odot (\mathbf{1} - \mathcal{M}_m). \tag{9}$$

Object-related and unrelated influences at step m are

$$\mathtt{I}^o_m = \sum_{s=1}^{S} \|\mathbf{g}^v_{ms}\|_1 \, \mathcal{M}_{ms}, \;\; \mathtt{I}^u_m = \sum_{s=1}^{S} \|\mathbf{g}^v_{ms}\|_1 \, (1 - \mathcal{M}_{ms}). \tag{10}$$

Masking and grouping are applied only during noun prediction (mitigate co-occurrence hallucination from object pairs). For non-noun steps, all elements in $\mathcal{M}_m$ are set to 0, yielding an empty $\mathbf{t}^o$.

Anchor-specific Influence-weighted Decoding. Let $\mathbf{z}^o_m = \pi_{\theta^\star}(\mathbf{t}^o, \mathbf{t}^p, \mathbf{y}_{<m})$ the *anchor-specific* negative logits and $\mathbf{z}^\star_m = \pi_{\theta^\star}(\mathbf{t}^v, \mathbf{t}^p, \mathbf{y}_{<m})$ be the *original* logits. We adjust logits by

$$\hat{\mathbf{z}}_m = (1 + \alpha_m)\, \mathbf{z}^\star_m - \alpha_m \, \mathbf{z}^o_m, \tag{11}$$

with $\alpha_m \geq 0$. In the probability space, moving along $\mathbf{z}^\star_m - \mathbf{z}^o_m$ increases the KL divergence $D_{\mathrm{KL}}(\sigma(\mathbf{z}^\star_m) \| \sigma(\mathbf{z}^o_m))$ (see Sec. 2 in supplementary). The original logits distribution $\mathbf{z}^\star_m$ can be viewed as $\pi_{\theta^\star}(\mathbf{t}^u, \mathbf{t}^o, \mathbf{t}^p, \mathbf{y}_{<m})$, i.e., a joint distribution that additionally depends on $\mathbf{t}^u$ compare to $\mathbf{z}^o_m$. Increasing the KL divergence therefore emphasizes the contribution of tokens $\mathbf{t}^u$, which are unrelated to previous mentioned objects, thereby mitigating co-occurrence bias in noun prediction. For non-noun steps, $\mathbf{t}^u$ coincides with $\mathbf{t}^v$, meaning that all visual tokens are emphasized. This adjustment helps reduce text–visual hallucination.

When analyzing token influence of $\hat{\mathbf{z}}_m$ in Eq. 11, the chain rule shows that $\mathbf{t}^u$ occur only in the original logits $\mathbf{z}^\star_m$ and are amplified by $(1 + \alpha_m)$, whereas other inputs ($\mathbf{t}^o$, $\mathbf{t}^p$, $\mathbf{y}_{<m}$) also contribute to $\mathbf{z}^o_m$ and therefore undergo smaller influence changes. Let $\tilde{\mathtt{I}}^o_m, \tilde{\mathtt{I}}^p_m, \tilde{\mathtt{I}}^y_m$ denote group-level influences computed on the negative branch $\mathbf{z}^o_m$ (analogous to (7)). We then choose α_m so that the influence of $\mathbf{t}^u$ matches the *dominant text* level, $\mathtt{I}^t_m := \max(\mathtt{I}^p_m, \mathtt{I}^y_m)$. Aligning $\mathbf{t}^u$ influence with the question prompt $\mathtt{I}^p_m$ is crucial for visually grounded responses, while balancing with previous outputs $\mathtt{I}^y_m$ prevents visual forgetting.

$$\alpha_m = \frac{\mathtt{I}^t_m - \mathtt{I}^v_m}{\mathtt{I}^v_m - \tilde{\mathtt{I}}^o_m + \tilde{\mathtt{I}}^t_m - \mathtt{I}^t_m}, \tilde{\mathtt{I}}^t_m = \begin{cases} \tilde{\mathtt{I}}^p_m & \text{if } \mathtt{I}^p_m \geq \mathtt{I}^y_m \\ \tilde{\mathtt{I}}^y_m & \text{otherwise} \end{cases} \tag{12}$$

Unlike existing decoding methods [12, 24, 54], which rely on adaptive plausibility constraints (e.g., prediction confidence) and require experimental tuning to determine optimal thresholds, our approach explicitly enforces non-negativity on the influence of object-related visual and prompt tokens. This corresponds to the following upper-bound condition:

$$0 \leq \alpha_m \leq \min\left\{ \frac{\mathtt{I}^o_m}{\tilde{\mathtt{I}}^o_m - \mathtt{I}^o_m}, \; \frac{\mathtt{I}^p_m}{\tilde{\mathtt{I}}^p_m - \mathtt{I}^p_m} \right\}. \tag{13}$$

Sample-dependent early stopping. Additionally, since hallucinations are more likely in long generations [35, 54], we introduce a sample-dependent stopping criterion based on visual influence. Specifically, if the visual influence ratio r^v_m of the token following the end-of-sequence (EOS) falls

below a threshold ϵ,

$$r_m^v := \frac{\mathtt{I}_m^v}{\mathtt{I}_m^v + \mathtt{I}_m^p + \mathtt{I}_m^y} < \epsilon \quad \text{and} \quad y_{m-1} = \text{EOS}. \quad (14)$$

Early stopping is triggered to prevent further output generation with minimal visual grounding.

4. Experiments

The proposed method is evaluated for both the open-ended generative tasks and the discriminative tasks. We use Amber [43], MSCOCO [28] and LLaVa-QA90 [30] datasets for the generative task, and Amber [43] and POPE [27] datasets on the discriminative tasks.

4.1. Evaluation Metrics

For generative image captioning, we focus on object hallucination and follow [10] report the Caption Hallucination Assessment with Image Relevance (CHAIR) [37] score, which includes sentence-level (hal,C_S) and instance-level (cha,C_I) percentages, instance-level recall (R,cov), and the average generated length (Len) [2], as well as co-occurrence object hallucination (cog) and the overall $score$ as suggested by [43]. For generative VQA, follow [17, 24] GPT-4V [1] is used to score both accuracy (Acc) and detailedness (Det) on a scale of 10. For discriminative tasks, hallucination manifests as a 'yes/no' misclassification we report both accuracy and F1 score.

4.2. Implementation Details

The maximum output length is set to 256 across all models, with other model parameters kept at their defaults. Gradients are obtained via PyTorch's torch.autograd.grad on the input tokens. Noun tokens are identified using the spaCy library via its `en_core_web_sm` model. To prevent excessive modifications, we set the maximum α_m to 5 for discriminative tasks and 3 for generative tasks. We empirically set the early stopping thresholds ϵ as follows: LLaVA-v1.5 and LLaVA-v1.6: 7%, InstrucBLIP: 25%, mPLUG-Owl2: 2.5%, and InternVL2: 10%. All experiments are performed on an NVIDIA A40 GPU with batch size of 1. Unless otherwise specified, we use greedy sampling [14].

4.3. Results on Open-ended Generation

In this section, we compare against SOTA alignment-based method RLAIF-V and contrastive decoding methods VCD, M3ID, and AVISC, on the AMBER and MSCOCO datasets, as presented in Tab. 1, Tab. 2. Additionally, we evaluate our method against VCD on the LLaVA-QA90 dataset, presented in Tab. 3. Our method outperforms most existing approaches across various baseline models and datasets, highlighting its robustness and reliable performance across different data types and model architectures. Specifically, we surpass image-level contrastive decoding methods like VCD and M3ID, demonstrating its effectiveness in operating at the token level and adapting to individual samples. Furthermore, compared to the token-level AVISC, our method excels, likely due to its object awareness and adaptability to fluctuating bias levels. The results further demonstrate that our method effectively mitigates hallucinations while preserving information.

Table 1. Results on the AMBER Dataset.

Method		Generative Task				Discriminative Task				Score
		cha ↓	cov ↑	hal ↓	cog ↓	acc ↑	P ↑	R ↑	F1 ↑	↑
LLaVA v1.5	base	7.8	**51.0**	36.4	4.2	72.0	**93.2**	62.4	74.7	83.5
	RLAIFv 47	6.6	49.7	32.0	2.9	76.7	92.0	78.1	84.5	89.0
	VCD 24	6.7	46.4	32.6	3.5	71.3	91.1	62.3	74.3	83.8
	M3ID 12	6.2	50.5	29.3	2.8	72.4	91.8	64.1	75.5	84.7
	AVISC 44	6.5	50.2	34.8	2.7	73.8	89.7	68.4	77.6	85.5
	Ours	**5.6**	**51.0**	**24.3**	**1.8**	**80.3**	82.9	**89.3**	**86.0**	**90.2**
Instruct BLIP	base	8.8	**52.2**	38.2	4.4	76.5	84.5	79.0	81.7	86.5
	RLAIFv 47	7.6	47.7	29.9	2.8	76.5	84.5	79.0	81.7	87.1
	VCD 24	7.9	49.7	36.7	3.7	75.9	83.5	79.3	81.3	86.7
	M3ID 12	7.3	49.2	33.8	3.7	75.8	84.4	77.9	81.0	86.9
	AVISC 44	7.1	48.8	34.4	4.3	75.9	83.4	**79.5**	81.4	87.2
	Ours	**6.0**	49.4	**26.6**	**2.4**	**78.1**	**88.8**	76.6	**82.2**	**88.1**
mPLUG Owl2	base	10.6	52.0	39.9	4.5	75.6	**95.0**	66.9	78.5	84.0
	RLAIFv 47	7.8	50.5	35.7	4.1	81.2	90.8	79.7	84.9	88.6
	VCD 24	8.0	51.3	35.3	4.1	75.6	83.5	78.8	81.1	86.6
	M3ID 12	7.8	51.7	34.9	4.1	75.9	83.5	79.3	81.3	86.8
	AVISC 44	10.9	50.5	35.5	4.4	82.1	90.7	81.4	85.8	87.5
	Ours	**7.5**	**53.6**	**34.7**	**4.0**	**82.1**	87.0	**86.2**	**86.6**	**89.6**
LLaVA v1.6	base	9.9	56.7	47.4	4.3	80.3	82.9	**89.3**	86.0	88.5
	RLAIFv	9.0	53.6	46.1	3.42	80.8	83.9	88.9	86.3	88.6
	VCD	9.5	52.7	46.3	3.78	79.9	83.1	87.6	85.4	88.0
	M3ID	9.2	50.1	45.3	3.3	80.4	83.2	88.8	85.9	88.4
	AVISC	9.2	50.7	47.5	3.2	80.6	83.5	88.2	85.8	88.3
	Ours	**8.7**	**58.3**	**43.8**	**2.5**	**81.2**	**85.2**	88.8	**87.0**	**89.2**
Qwen2 VL	base	6.4	70.4	54.8	5.9	82.9	**91.6**	82.2	86.6	90.1
	RLAIFv	5.8	69.4	54.1	5.5	83.5	91.2	82.6	86.7	90.4
	VCD	6.5	69.1	53.7	5.3	82.7	90.9	82.3	86.4	90.0
	M3ID	6.3	68.8	53.5	5.1	83.0	91.0	82.8	86.7	90.2
	AVISC	6.3	69.0	53.9	5.0	82.8	91.1	82.5	86.6	90.1
	Ours	**4.9**	**71.8**	**44.7**	**3.7**	**84.4**	88.1	**89.2**	**87.1**	**91.1**
Intern VL2	base	8.1	69.6	59.0	5.2	84.0	87.3	**88.8**	88.0	90.0
	RLAIFv	8.0	68.4	59.3	4.9	84.2	87.7	88.5	88.1	90.1
	VCD	8.5	68.7	58.6	5.0	82.9	87.0	88.4	87.7	89.6
	M3ID	8.4	69.2	58.9	5.4	83.7	86.8	88.4	87.6	89.6
	AVISC	8.4	68.9	59.1	4.8	84.0	87.7	86.8	87.2	89.4
	Ours	**7.9**	**69.8**	**57.8**	**3.7**	**84.7**	**88.2**	**88.8**	**88.5**	**90.3**

Hallucination Mitigation. Our approach reduces hallucination by up to 33% at sentence-level (hal in Tab. 1 and C_S in Tab. 2) and 32% at instance-level (cha in Tab. 1 and C_I in Tab. 2), demonstrating its effectiveness in mitigating overall hallucinations. It also effectively mitigates co-occurrence hallucinations, with reductions of up to 57% for cog in Tab. 1. Accuracy gains of up to 92% (Tab. 3) further demonstrate that our model improves alignment with the input image, highlighting its ability to jointly address text–visual and co-occurrence biases (Sec. 5).

Information Preservation. Our method also enhances information preservation, with recall (cov in Tab. 1 and R in

[2] Since shorter outputs can trivially lower CHAIR scores at the expense of informativeness.

Table 2. Open-ended Generation Results Using the CHAIR Metric on the MSCOCO Subset Following [10].

Models	Metrics	Baseline	VCD	M3ID	AVISC	Ours
LLaVA-v1.5	$C_S \downarrow$	48.8	44.8	44.5	46.4	**41.0**
	$C_I \downarrow$	13.4	12.8	12.1	13.4	**10.9**
	$R \uparrow$	**78.6**	76.8	77.0	76.3	77.3
	$Len \uparrow$	99.8	89.8	85.1	90.5	85.0
InstructBLIP	$C_S \downarrow$	57.8	63.4	57.3	58.9	**47.4**
	$C_I \downarrow$	16.5	19.6	16.1	17.8	**13.4**
	$R \uparrow$	**73.6**	71.2	72.5	70.6	72.3
	$Len \uparrow$	101.3	95.5	100.1	99.6	93.9
mPLUG-Owl2	$C_S \downarrow$	59.2	52.7	52.4	58.3	**45.0**
	$C_I \downarrow$	17.6	16.1	15.8	17.5	**12.4**
	$R \uparrow$	**75.8**	73.2	72.7	75.6	74.9
	$Len \uparrow$	105.3	93.6	92.6	99.5	83.5

Tab. 2) dropping by an average of only 1.1%, compared to an average drop of 3.2% in other methods, and even increasing by 3.1% when using the baseline mPLUG-Owl2 on the AMBER dataset. Higher recall indicates that our model retrieves a broader range of objects from visual inputs. Additionally, results in Tab. 3 an increase of up to 45% in detailedness (Det), further demonstrating our method's effectiveness in retrieving all relevant visual details and mitigating visual forgetting.

Table 3. Results on LLaVA-QA90, settings following [24].

Method	LLaVA-v1.5		IntructBLIP		mPLUG-Owl2	
	Acc↑	Det↑	Acc↑	Det↑	Acc↑	Det↑
base	3.23	3.54	3.84	4.07	4.07	4.33
VCD	4.15	3.85	4.23	4.69	4.52	4.64
M3ID	4.57	3.96	4.67	4.61	4.44	5.18
AVISC	4.88	3.87	4.32	4.27	4.75	5.12
RLAIF-V	5.79	4.74	5.27	4.62	5.03	5.33
Ours	**6.20**	**5.13**	**6.28**	**4.77**	**6.69**	**6.28**

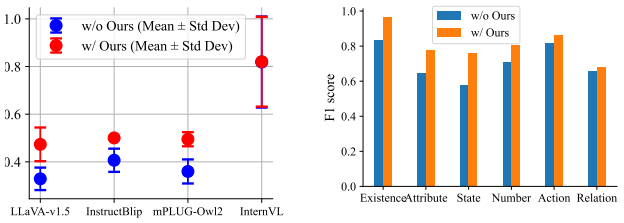
4.4. Results on Discriminative Task

We next evaluate our method on discriminative tasks using AMBER (discriminative VQA) and POPE (existence VQA), with results shown in Tab. 1 and Tab. 4. Our approach achieves a better balance between precision and recall, yielding consistently higher F1 scores and improved accuracy. Notably, unlike competing methods that degrade Intern-VL2, ours preserves its performance via bias awareness. Category-wise analysis further shows heterogeneous gains, with improvements varying across question types.

Variation in Improvement Across Question Categories. Fig. 3b presents F1 scores across various question categories using LLaVA-v1.5 [29]. Our method improves performance across all categories, with the largest gains in existence, attributes, and state—categories strongly tied to visual cues, benefiting from enhanced visual token influence.

Table 4. Results on POPE in MSCOCO Adversarial Setting.

Method	LLaVA v1.5		Instruct BLIP		mPLUG Owl2		Intern VL2	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
base	80.9	81.6	79.8	81.4	72.5	77.5	**85.8**	**85.0**
VCD	80.9	81.3	79.6	79.5	74.2	78.8	83.2	82.2
M3ID	81.7	81.8	81.0	81.6	75.6	79.1	83.5	82.1
AVISC	81.2	81.6	81.8	81.9	80.9	79.7	85.3	84.6
Woodpecker	80.5	80.6	79.0	78.6	77.5	76.9	85.7	84.8
Ours	**83.5**	**82.1**	**82.5**	**82.1**	**84.2**	**83.7**	**85.8**	**85.0**



(a) Visual Influence Ratio (b) F1 across Question Categories

Figure 3. (a) Visual influence ratios across the POPE dataset, illustrating variation across MLLMs. Our method successfully increases the visual influence ratio when it falls below 50%. (b) F1 scores for the AMBER discriminative task using LLaVA-v1.5 are consistently improved by our method, with particularly notable gains in the existence and state categories.

Variation in Improvement Across MLLMs. Our method achieves the most significant improvement on mPLUG-Owl2 (Tab. 4) and on LLaVA-v1.5 (Tab. 1), Consistent performance gains are observed across modern MLLMs (LLaVA-v1.6, Qwen2-VL, and InternVL2). Compared with static heuristics (e.g., VCD), which may degrade performance due to overcorrection, our method maintains stable improvements across models. We further analyze and find that performance variation is correlated with the baseline visual influence ratio. Fig. 3a presents the visual influence ratios in object existence VQA, showing that LLaVA-v1.5 exhibits the lowest visual contribution, followed by mPLUG-Owl2. This lower baseline visual influence ratio allows our method to make more impactful adjustments. In contrast, InternVL2 has an original visual influence ratio exceeding 50%, resulting in minimal improvement when our method is applied. The strong performance of InternVL2 can be attributed to its original high visual influence ratio, further validating the motivation behind our approach.

5. Ablation Study

In this section, we first analyze text–visual and co-occurrence biases, then evaluate the contributions of our proposed components. We also detail the gradient-computation methods and norm selection. Additional ablation studies and hyperparameter settings are provided in the Supplementary Material.

Text-Visual Bias Analysis. Fig. 3a shows that with the exception of InternVL2, MLLMs (LLaVA-v1.5, InstructBLIP, and mPLUG-Owl2) rely more on text prompt than on visual input. Likely due to MLLMs' training process, this tendency is common in MLLMs, where multimodal features are aligned with language tokens after extensive text-based pre-training, causing language components to dom-

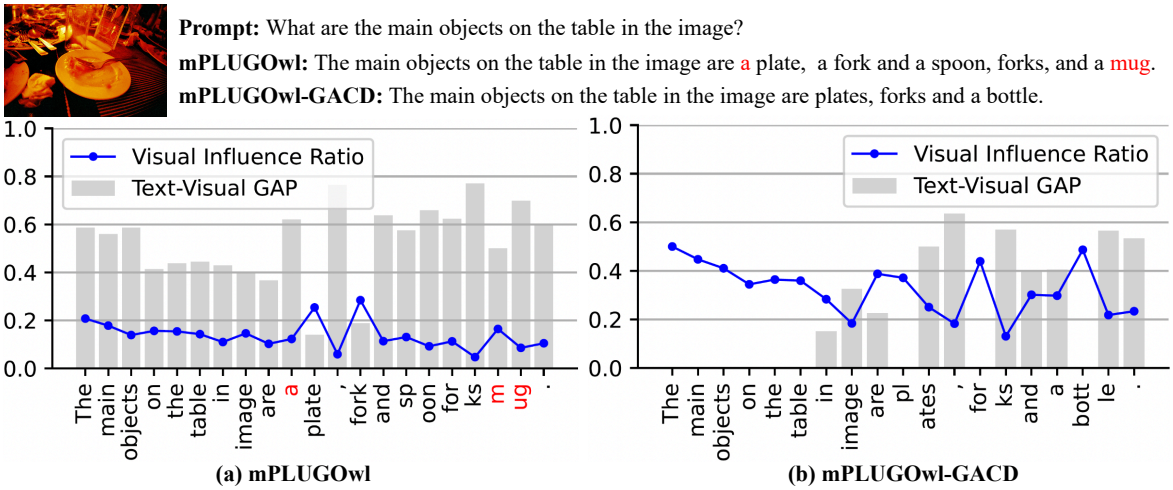


Figure 4. Comparison of visual influence ratios r_m^v and Text-Visual GAP, with and without our GACD. (a) Without GACD, mPLUG-Owl2 shows a low initial visual influence ratio, punctuation marks and suffixes naturally have low visual influence, while objects start with higher influence that declines as the sequence grows. Hallucinations tend to occur when the visual ratio is low. The text-visual gap confirms that text dominates the influence on predictions. (b) With GACD, the visual influence ratio increases overall and mitigates the decrease over the sequence length. The text tokens only domain influence in predictions less related to the visual, reducing hallucination.

inate decision-making. GACD effectively increases overall visual influence to match that of object-present question prompts in POPE (Fig. 3a). In the open-ended generation task, we further observe the visual influence ratio r_m^v and the Text-Visual GAP, defined as $\max(\max(r_m^p, r_m^y) - r_m^v, 0)$, the difference between the text influence ratio and the visual influence ratio when text influence is dominant [3]. Observations in Fig. 4 also highlight the text-dominant influence typical of MLLMs. GACD counteracts this by boosting the influence of visual tokens when aligning them with prompts and previous outputs, leading to higher prediction confidence and a reduction in hallucinations (Fig. 1 in supplementary). Additionally, the nature of the output token affects the visual influence ratio. For instance, punctuation marks or suffixes tend to have a lower visual influence ratio. This is intuitive, as these tokens rely more on linguistic context and are less dependent on visual information. This observation highlights the value of our GACD framework delivering sample-dependent, token-specific hallucination mitigation.

Co-occurrence Bias Analysis. Fig. 5a shows an example where mPLUG-Owl2 incorrectly predicts 'dining table' due to the presence of a 'chair'. In Fig. 5b, the influence of individual visual tokens on hallucinated prediction I_{ms}('table') and I_{ms}('chair') shows that they share the same most influential visual token: $s = 24$. These visualizations indicate that the influence distribution over tokens is typically sharply peaked, so selecting the single most influential token gives a clean and interpretable attribution signal while reducing noise from low-influence tokens. We further collected 100 chair-only and 100 table-only images from MSCOCO evaluation dataset [28].

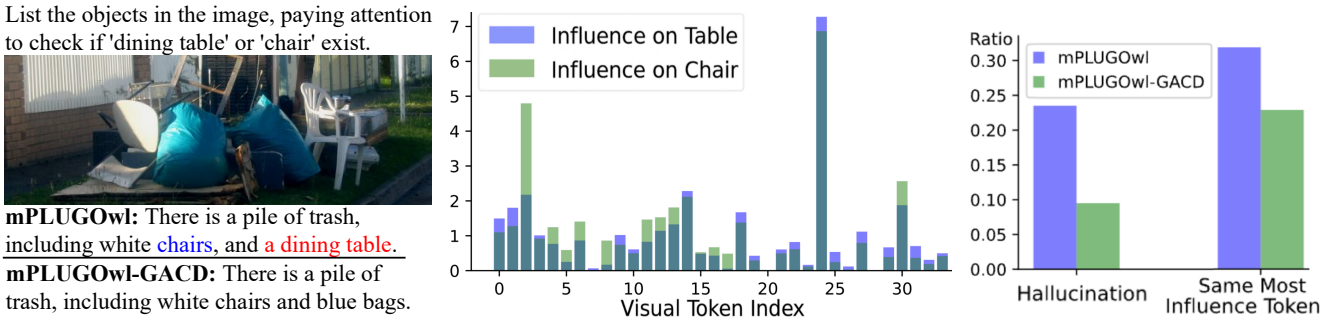
Results in Fig. 5c show that when only a 'chair' or 'table' exists in the image, the other object is hallucinated in 23.5% of cases, with 31.9% sharing the same most influential token, indicating that the 'Same Most Influential Token' phenomenon is common in co-occurrence hallucinations. Our GACD effectively reduces such hallucination where both 'table' or 'chair' are predicted in single-object images.

Table 5. Component Analysis Using the CHAIR Metric.

Models	VA / CR / ES		✓	✓ ✓	✓ ✓ ✓
LLaVA-v1.5	$C_S \downarrow$	48.8	46.4	46.2	**41.0**
	$C_I \downarrow$	13.4	11.6	11.3	**10.9**
	$R \uparrow$	78.6	79.0	**79.4**	77.3
	$Len \uparrow$	**99.8**	95.6	95.5	85.0
InstructBLIP	$C_S \downarrow$	57.8	53.6	53.2	**47.4**
	$C_I \downarrow$	16.5	15.1	14.0	**13.4**
	$R \uparrow$	73.6	**75.3**	74.6	72.3
	$Len \uparrow$	101.3	**108.4**	105.7	93.9
mPLUG-Owl2	$C_S \downarrow$	59.2	52.6	52.3	**45.0**
	$C_I \downarrow$	17.6	14.4	14.2	**12.4**
	$R \uparrow$	75.8	**78.2**	78.0	74.9
	$Len \uparrow$	**105.3**	95.6	95.5	83.5

Component Analysis. To assess the effectiveness of each component in our proposed method, we conducted the following variants: 1) Visual Amplification (VA) only: visual amplification is applied to all visual tokens ($\mathbf{t}^v$), including during noun predictions. 2) Co-occurrence Hallucination Reduction (CR): object-related visual tokens are detected, and $\mathbf{t}^u$ is amplified during noun predictions. 3) Our full model, with early stopping (ES). Tab. 5 demonstrates that each component of our method contributes to the overall performance. VA significantly reduces hallucinations while improving object recall. CR further mitigates co-occurrence

[3] r_m^p and r_m^y are derived in the same manner as r_m^v in (14).



(a) Co-occurrence Hallucination (b) Individual Visual Token Influence (c) Overall Statistics

Figure 5. Co-occurrence hallucination of 'table' in the presence of 'chair'. (a) Comparison of outputs with and without GACD. (b) Visualization of individual visual token influence indicates that the visual token with index 24, which has the highest influence on the hallucinated 'table', also holds the highest influence on 'chair'. (c) Summary statistics for 100 chair-only and 100 table-only images, showing the hallucination rate and the percentage of cases where both objects share the same most influential visual token (as illustrated in b). GACD effectively reduces both metrics.

bias, a residual form of the text-visual bias addressed by VA, resulting in additional hallucination reduction. Both VA and CR achieve these gains without introducing trade-offs. When necessary, the ES mechanism shortens outputs to effectively reduce hallucinations, with only a slight recall trade-off.

Gradient Computation. Our method obtains gradients directly through PyTorch's 'torch.autograd.grad' on input tokens, eliminating the need for manual derivations and enabling straightforward reproducibility. For comparison, we evaluate Integrated Gradients (IG) [11, 21, 31] using the SIG [11] implementation; results for this ablation on the POPE MSCOCO adversarial setting are shown in Tab. 6. 'IG' denotes the SIG-based results, while "Ours" refers to our direct gradient method. Both achieve comparable accuracy and F1-score, but the direct-gradient variant is substantially more efficient.

Table 6. Gradient Methods on the POPE MSCOCO Dataset(Adv.)

Methods		Accuracy	F1	Speed(ms)
MPLUG-Owl2	IG [11]	83.4	82.9	20335
	Ours	84.2	83.7	385

Norm Analysis. We also study the effect of norm selection on token influence, comparing L1 (Manhattan), L2 (Euclidean), and L∞ (infinity) norms. The L1 norm emphasizes individual token contributions, the L2 norm reflects overall influence, and the L∞ norm focuses on the strongest activation. As shown in Tab. 7, the L1 norm yields the best performance, supporting the intuition that it effectively captures influence magnitude across tokens and channels.

Table 7. Norm Strategies on the POPE MSCOCO Dataset(Adv.)

Norm	LLaVA-v1.5		InstructBLIP		mPLUG-Owl2	
	Acc	F1	Acc	F1	Acc	F1
L1	**83.5**	**82.1**	**82.5**	**82.1**	**84.2**	**83.7**
L2	83.2	81.9	79.5	79.6	83.2	82.9
L$_\infty$	83.4	82.0	82.1	81.8	80.8	80.6

Cost and Runtime. We further analyze computational cost and runtime. The visual encoder is executed only once, and the second pass operates on a small set of tokens. On the POPE dataset, this yields a 101.44% increase in average, comparable to decoding-based methods (e.g., VCD) that also require additional guidance computation.

Table 8. Runtime Comparison on POPE MSCOCO Dataset(Adv.)

LLaVA-v1.5	TFLOPs	Runtime (ms)	Increase
base	9.68	191	–
VCD	19.37	383	+100.10%
Ours	19.49	385	+101.44%

6. Conclusion

In conclusion, we introduce a gradient-based self-reflection method to estimate token influence and quantitatively estimate bias severity. This estimation enables the identification of object-related visual tokens, which are then integrated into an influence-aware constrained decoding framework. This framework effectively mitigates both text-visual and co-occurrence biases, reducing hallucinations. Our method operates without requiring additional resources such as costly fine-tuning, extra models, or data statistics. Furthermore, to reduce text-visual bias in long-generated sequences, we propose a sample-dependent stopping criterion based on the proposed visual influence.

Limitations. Our method is limited to white-box MLLMs, as it requires access to gradients. Its effectiveness depends on the baseline MLLM's original visual influence ratio, and the importance of visual information. For instance, existence questions are primarily rely on visual, whereas relational questions are less direct and require inference beyond visual. As a post-processing technique, our method does not involve model training. In future work, we aim to explore how insights from GACD can guide and improve training strategies for enhanced visual perception in MLLMs.

Acknowledgments

This work was conducted during an internship at NVIDIA. Hongdong Li is also partially supported by an ARC Discovery Grant DP220100800.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1

[3] Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mocha: Multi-objective reinforcement mitigating caption hallucinations. *arXiv preprint arXiv:2312.03631*, 2023. 1

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[5] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023. 1

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1

[7] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024. 1, 2

[8] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023. 2

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[10] Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024. 1, 2, 5, 6

[11] Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. *arXiv preprint arXiv:2305.15853*, 2023. 8

[12] 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. *arXiv preprint arXiv:2403.14003*, 2024. 1, 2, 4, 5

[13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 2

[14] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 5

[15] Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*, 2024. 2, 3

[16] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 2

[17] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024. 5

[18] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 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, 2024. 1, 2

[19] Zhehan Kan, Ce Zhang, Zihan Liao, Yapeng Tian, Wenming Yang, Junyuan Xiao, Xu Li, Dongmei Jiang, Yaowei Wang, and Qingmin Liao. Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms, 2024. 1, 2

[20] Cheongwoong Kang and Jaesik Choi. Impact of co-occurrence on factual knowledge of large language models. *arXiv preprint arXiv:2310.08256*, 2023. 1

[21] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021. 8

[22] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9012–9020, 2019. 2, 3

[23] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11082–11092, 2024. 1, 2

[24] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 1, 2, 4, 5, 6

[25] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023. 2

[26] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*, 2022. 3

[27] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1, 2, 5

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5, 7

[29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 6

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 5

[31] Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pages 14485–14508. PMLR, 2022. 8

[32] Avshalom Manevich and Reut Tsarfaty. Mitigating hallucinations in large vision-language models (lvlms) via language-contrastive decoding (lcd), 2024. 2

[33] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*, 2023. 2

[34] Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2408.13906*, 2024. 1

[35] Shangpin Peng, Senqiao Yang, Li Jiang, and Zhuotao Tian. Mitigating object hallucinations via sentence-level early intervention. *arXiv preprint arXiv:2507.12455*, 2025. 2, 4

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[37] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018. 1, 5

[38] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 3

[39] Michael Spivak. Calculus. houston, tx: Publish or perish, 1980. 3

[40] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 1

[41] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[42] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024. 2

[43] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 2, 5

[44] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024. 1, 2, 5

[45] Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2402.09801*, 2024. 1, 2

[46] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 1

[47] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 5

[48] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024. 1, 2

[49] Fatemeh Pesaran Zadeh, Yoojin Oh, and Gunhee Kim. Lpoi: Listwise preference optimization for vision language models. *arXiv preprint arXiv:2505.21061*, 2025. 2

[50] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. Halle-control: Controlling object hallucination in large multimodal models, 2024. 2

[51] Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv e-prints*, pages arXiv–2406, 2024. 2

[52] Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing large visual language models. *arXiv preprint arXiv:2403.05262*, 2024. 2

[53] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*, 2024. 2

[54] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 1, 2, 4

[55] Yibo Zhou, Hai-Miao Hu, Jinzuo Yu, Zhenbo Xu, Weiqing Lu, and Yuran Cao. A solution to co-occurrence bias: Attributes disentanglement via mutual information minimization for pedestrian attribute recognition. *arXiv preprint arXiv:2307.15252*, 2023. 1