Causally-Grounded Dual-Path Attention Intervention for Object Hallucination Mitigation in LVLMs

Liu Yu[1,2*], Zhonghao Chen[1], Ping Kuang[1†], Zhikun Feng[1], Fan Zhou[1], Lan Wang[1], Gillian Dobbie[2]

[1]University of Electronic Science and Technology of China,
[2]University of Auckland

liu.yu@std.uestc.edu.cn, 202321090211@std.uestc.edu.cn, kuangping@uestc.edu.cn, 202411090917@uestc.edu.cn,
fan.zhou@uestc.edu.cn, 202421090210@std.uestc.edu.cn, g.dobbie@auckland.ac.nz

Abstract

Object hallucination remains a critical challenge in Large Vision-Language Models (LVLMs), where models generate content inconsistent with visual inputs. Existing language-decoder based mitigation approaches often regulate visual or textual attention independently, overlooking their interaction as two key causal factors. To address this, we propose **Owl** (Bi-m**O**dal attention re**W**eighting for **L**ayer-wise hallucination mitigation), a causally-grounded framework that models hallucination process via a structural causal graph, treating decomposed visual and textual attentions as mediators. We introduce VTACR (Visual-to-Textual Attention Contribution Ratio), a novel metric that quantifies the modality contribution imbalance during decoding. Our analysis reveals that hallucinations frequently occur in low-VTACR scenarios, where textual priors dominate and visual grounding is weakened. To mitigate this, we design a fine-grained attention intervention mechanism that dynamically adjusts token- and layer-wise attention guided by VTACR signals. Finally, we propose a dual-path contrastive decoding strategy: one path emphasizes visually grounded predictions, while the other amplifies hallucinated ones – letting visual truth shine and hallucination collapse. Experimental results on the POPE and CHAIR benchmarks show that Owl achieves significant hallucination reduction, setting a new SOTA in faithfulness while preserving vision-language understanding capability. Our code is available at https://github.com/CikZ2023/OWL

1 Introduction

Large Vision-Language Models (LVLMs) such as MiniGPT-4 (Zhu et al. 2023), LLaVA (Liu et al. 2023) and Shikra (Chen et al. 2023), have achieved impressive progress in image-based text generation (Li et al. 2023b, 2022), allowing a wide range of applications, from visual question answering (Kim et al. 2025) to open-ended image description (Liu et al. 2024b; Yu et al. 2025c). Despite their success, these models remain vulnerable to a persistent issue – *object hallucination* – that generates mentions of objects
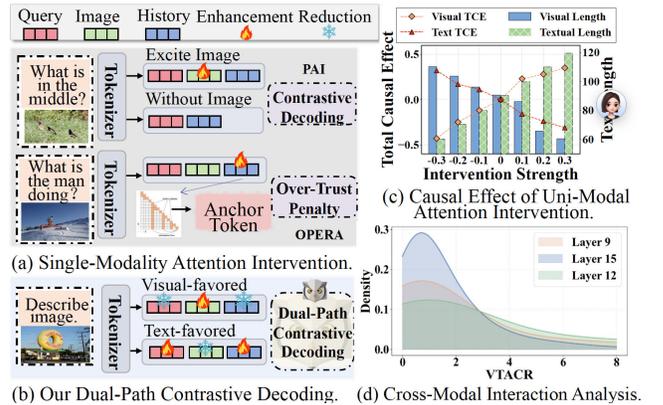
Figure 1: Motivation of our work. (a) Existing methods manipulate attention in a single modality (visual or text). (b) We contrast the visual-favored path and text-favored path based on the VTACR-guided attention calibration. (c) Increasing visual attention improves causal effect but shortens output, while increasing textual attention has the opposite impact. (d) Hallucinated tokens typically show lower VTACR, indicating a skewed visual-to-textual modality reliance.

not present in the image. Such hallucinations not only undermine the trustworthiness of LVLMs but also pose serious risks in safety-critical domains such as medical imaging (Hu et al. 2024), robotic navigation (Lange et al. 2025), etc.

To mitigate this issue, existing approaches span several directions. Early efforts align LVLMs with human preferences (Sun et al. 2023; Gunjal, Yin, and Bas 2024) through reinforcement learning or feedback-based fine-tuning, which improves consistency but often requires costly annotations. Others adopt post-processing strategies (Zhou et al. 2023b; Deng, Chen, and Hooi 2024; Yin et al. 2024) using external modules to detect or revise hallucinated entities after generation. More recently, decoding optimizations have gained traction: Some (Leng et al. 2024) perturb visual inputs to reveal unstable predictions, while others (Liu, Zheng, and Chen 2024; Huang et al. 2024b) manipulate attention weights to boost visual grounding or suppress over-reliance on previous tokens. However, as shown in Figure 1(a), language-decoder attention-based methods tend to

act on either the visual path – by enhancing attention to image tokens – or the textual path – by diminishing influence from autoregressive history. This uni-modal design overlooks the attention imbalance between the visual and textual modalities, which often lie at the heart of hallucination. To better understand this issue, we decompose the attentions in language-decoder into visual and textual aspects, and analyze both uni-modal and their interaction. As shown in Figure 1(c), we observe that: (1) solely enhancing visual attention consistently reduces hallucinations (indicated by increasing Total Causal Effect, TCE) but leads to shorter outputs; (2) in contrast, solely increasing textual attention (including query and history) expands output length but aggravates hallucinations. This tradeoff suggests the need to balance their reliance rather than treating them independently. To measure their interaction, we introduce a new metric – visual-to-textual attention contribution ratio (VTACR) – to quantify the relative hallucination contribution of visual versus textual signals for the current token during generation:

- **Visual Token Attention Contribution**:

$$\nu^{(\ell)} = \frac{1}{N|\mathcal{V}|} \sum_{j \in \mathcal{V}} \sum_{i=1}^{N} \mathbf{A}_{i,j}^{(\ell)} \quad (1)$$

where $\mathcal{V}$ is the set of indices for visual prefix tokens, and N is the number of attention heads. $\mathbf{A}_{i,j}^{(\ell)}$ indicates the visual attention weight of the current token in the i-th head and the ℓ-th layer, and $\nu^{(\ell)}$ is the average of the attention weight of the current token to the visual prefix tokens.

- **Text Token Attention Contribution**:

$$\tau^{(\ell)} = \frac{1}{N|\mathcal{T}|} \sum_{k \in \mathcal{T}} \sum_{i=1}^{N} \mathbf{A}_{i,k}^{(\ell)} \quad (2)$$

where $\mathcal{T}$ is indices set for query and history text prefix tokens, $\mathbf{A}_{i,k}^{(\ell)}$ indicates the text attention weight of the current token in i-th head, ℓ-th layer, and $\nu^{(\ell)}$ is the average of the attention weight of the current token to the text prefix tokens.

- **Layer-wise VTACR**:

$$\text{VTACR}^{(\ell)} = \frac{\nu^{(\ell)}}{\tau^{(\ell)}} \quad (3)$$

This ratio measures the relative contribution of visual tokens to text tokens in the ℓ-th layer. L is the total layers. As shown in Figure 1(d), hallucinated tokens tend to exhibit skewed VTACR values in LLaVA-1.5, revealing a tendency to over-rely on textual modality while neglecting visual grounding (consistent across layers/backbones). This modality imbalance motivates us to explicit decomposition of two contrasting attention pathways: a *vision-favored* path that reinforces grounded reasoning, and a *text-favored* path that tends to preserve hallucinated content. Such separation enables us to capture the asymmetric roles of each modality in hallucination formation and lays the foundation for a contrastive decoding mechanism.

To formalize this intuition, we construct a Structural Causal Model (SCM) in which visual and textual attention serve as *mediators* between inputs and outputs. Unlike traditional causal interventions on inputs or latent states (Huang et al. 2024a; Zhou et al. 2025), mediator-based intervention enables direct manipulation of attention weights while preserving input consistency. This provides a more interpretable and fine-grained view of how modality interactions drive hallucinations. We further leverage VTACR to guide token- and layer-wise attention adjustment. Rather than applying fixed scaling factors (as in previous work like PAI (Liu, Zheng, and Chen 2024)), by adjusting the token-by-token attention weights according to real-time modality contributions, especially in layers with pronounced asymmetry – our Owl not only corrects attention imbalance, but also improves generation quality as a side benefit, producing longer outputs with fewer repetitions and less overcorrection. Finally, we introduce a dual-path contrastive decoding strategy, incorporating both vision- and text-favored decoding paths. The contrast amplifies the distinction between hallucinated and faithful tokens, effectively mitigating hallucinations. Our contributions are four-fold:

- We introduce a novel metric VTACR to quantify cross-modal reliance during generation, and use it to guide fine-grained token- and layer-wise attention modulation.
- We formulate a SCM where visual and textual attention serve as mediators, enabling interpretable dual-modality interventions to analyze the hallucinations process.
- We propose a VTACR-guided dual-path contrastive decoding strategy that exaggerates modality bias – amplifying faithful generations while exposing hallucinated ones – thereby enabling effective suppression through contrast.
- Experiments on POPE/CHAIR show Owl achieves notable hallucination reduction, with 22.9% improvement on CHAIR, while evaluations on five VQA benchmarks confirm preserved vision-language understanding capability.

2   Related Work

Object Hallucination Mitigation in LVLMs.   Object hallucination in LVLMs often stems from over-reliance on spurious correlations – models tend to exploit shortcut patterns, such as object co-occurrence or prompt biases, learned from large-scale vision-language data. This results in fluent but visually unfaithful generations (Lyu et al. 2025; Zhou et al. 2024; Yu et al. 2025a). To address this, existing methods fall into three main paradigms: (1) Early *human preference alignment* approaches like LLaVA-RLHF (Sun et al. 2023) and instruction tuning fine-tune models to match human-preferred responses, (Bai et al. 2025). While helpful for fluency and helpfulness, these methods are costly and offer limited interpretability into hallucination origins. (2) *Post-processing* works like LURE (Zhou et al. 2023b), CGD (Deng, Chen, and Hooi 2024), and Woodpecker (Yin et al. 2024) detect hallucinations post-hoc via confidence scoring or visual grounding checks. These modular approaches offer flexibility but rely heavily on external cues and do not address the root cause. (3) Recent *decoding optimization* works proactively steer generation to reduce hallucinations. Contrastive decoding methods (e.g., VCD (Leng et al. 2024), HIO (Lyu et al. 2024)) amplify hallucinated signals to isolate faithful ones. Others manipulate attention to strengthen visual grounding (Liu, Zheng, and Chen 2024) or suppress misleading text (Huang et al. 2024b). Token-level

signals like VAR (Jiang et al. 2024) and "attention sin" patterns (Zhang et al. 2024) enhance interpretability, yet most still intervene on a single modality, overlooking the joint causal role of visual and textual attention.

Causality in LVLMs. Causal inference (Pearl 2010; Neal 2020) offers a powerful lens to enhance interpretability and robustness in AI systems, particularly do-calculus or counterfactual simulations to uncover causal links in language generation (Zhang et al. 2025a), visual question answering (Zhang et al. 2025b), and fairness analysis (Yu et al. 2025b; Zhou et al. 2023a). These approaches typically intervene on inputs or latent features to disentangle causal effects from spurious correlations. Recent works use this tool for object hallucination: Huang et al. (2024a) intervenes on image/text inputs and embeddings to analyze hallucination triggers, while CausalMM (Zhou et al. 2025) perturbs attentions in vision and language decoder to probe modality priors. However, they often rely on coarse-grained manipulations or treat attention as a black box. In contrast, we model visual and textual attention as explicit mediators within a SCM, allowing fine-grained mediator interventions that directly adjust internal attention without altering the input, enabling interpretable causal analysis of modality influence.

3 Preliminary

Formulation of LVLM Generation. The LVLMs typically consist of three key components: a visual encoder, a cross-modal projector, and a language decoder. The visual encoder (e.g., ViT (Khan et al. 2022)) extracts a sequence of image features $\mathbf{X}_V = [x_{v_1}, \ldots, x_{v_N}]$, which are mapped into the text embedding space via a cross-modal projector (Alayrac et al. 2022; Li et al. 2023a; Liu et al. 2023). The projected visual tokens are then concatenated with the instruction text $\mathbf{X}_T = [x_{i_1}, \ldots, x_{i_M}]$ and historical textual input $\mathbf{X}_H = [x_{h_1}, \ldots, x_{h_L}]$, forming the input to the language decoder (e.g., LLaMA (Touvron et al. 2023)).

The decoder integrates the multi-modal inputs via multi-layer attention and produces contextualized hidden states. The hidden state $\mathbf{h}_t$ at a target position t (typically the last token in $\mathbf{X}_H$) is used to compute the token probability:

$$\mathbf{h}_t = \text{Decoder}_\theta(\mathbf{X}_V, \mathbf{X}_I, \mathbf{X}_H)[t], \tag{4}$$

$$P_\theta(y_t \mid \mathbf{X}_V, \mathbf{X}_T, \mathbf{X}_H) = \text{Softmax}(\mathbf{W}_o \cdot \mathbf{h}_t), \tag{5}$$

where $\mathbf{W}_o \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the output projection matrix and $|\mathcal{V}|$ denotes the vocabulary size. Autoregressive decoding continues until an end-of-sequence (EOS) token is produced, forming the final output $y_{1:T}$. The goal of object hallucination mitigation is to ensure that the generated object-level content aligns faithfully with the visual evidence in $\mathbf{X}_V$, such that the outputs are both semantically relevant and visually grounded. The overall framework is shown in Figure 3.

4 Methodology

Causal Modeling of Object Hallucination Process. We construct a SCM in Figure 2. The attention mechanisms in
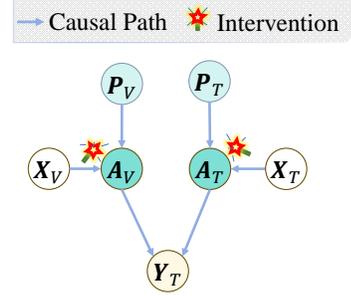

Figure 2: The SCM for analyzing the hallucination process. Visual input (X_V) and text input (X_T) affect the output (Y_T) via visual attention (A_V) and text attention (A_T). Visual priors (P_V) and language priors (P_T) confound the attention paths and may cause hallucinations. Interventions on A_V and A_T help estimate their causal impact.

LLaMA-style language decoders serve as a core computational unit, and we explicitly decouple attention into visual and textual attention components, which are influenced by both input modalities and modality-specific priors. Specifically, the SCM consists of image input X_V, text input X_T, priors P_V and P_T, visual and textual attention A_V and A_T, and final language output Y_T. The causal relations can be summarized as:

$$X_V \to A_V \to Y_T, \quad P_V \to A_V \to Y_T,$$
$$X_T \to A_T \to Y_T, \quad P_T \to A_T \to Y_T.$$

where these paths reflect how hallucinations may be introduced via unbalanced attention induced by modality-specific inputs and priors. Notably, since the priors P_V and P_T are unobservable or non-manipulable, we cannot intervene on them directly. However, as they influence the output exclusively through the *mediators* A_V and A_T, we target these attention modules for causal intervention. To this end, we apply soft interventions on A_V and A_T:

$$do(A_V = A_V^*), \quad do(A_T = A_T^*), \tag{6}$$

where A_V^* and A_T^* are calibrated attention weights obtained through our debiasing module. This mediator intervention conforms to the do-calculus formulation in causal inference (Pearl 2022), representing a hypothetical manipulation that isolates the effect of attention from upstream biases.

To evaluate the effect of such interventions, we adopt the Total Causal Effect (TCE) as our primary metric. TCE measures the average change in hallucination behavior caused by modifying the mediators, and is defined as:

$$\text{TCE} = \mathbb{E}_{x \sim X_{\text{test}}} \big[\Psi \big(Y_T(P, A), Y_T'(P, A^*) \big) \big] \tag{7}$$

$$\Psi \big(Y_T, Y_T' \big) = 2 \cdot \mathbb{I} \big(H(Y_T) > H(Y_T') \big) - 1 \tag{8}$$

where Ψ measures the causal effect metric between the distributions before and after the intervention. $H(\cdot)$ denotes the hallucination evaluation benchmark CHAIR in Sec. 5, and Y_T' is the output under intervened attention. $I(\cdot)$ denotes the indicator function. If $H(Y_T) > H(Y_T')$ holds, the indicator result is 1, indicating that the hallucination is reduced; otherwise, it is 0, indicating that the hallucination is not reduced.
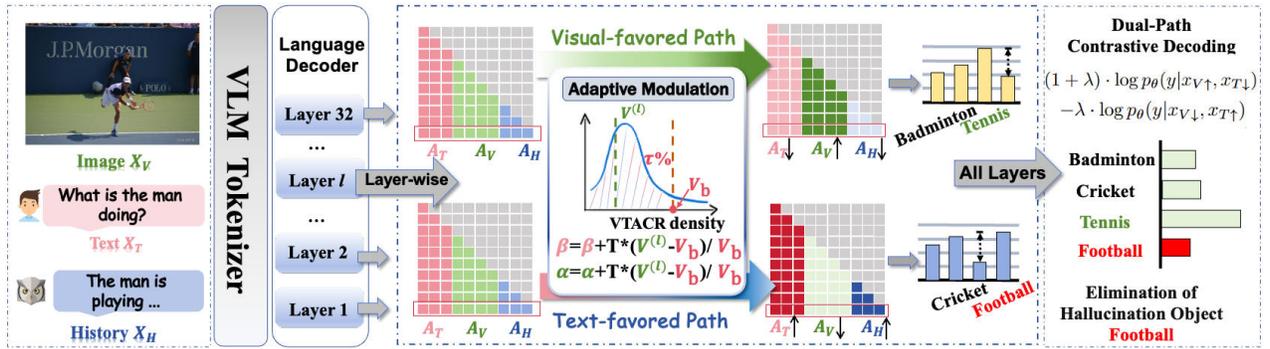
Figure 3: The overall framework of **Owl**. Given image, text, and generation history, Owl performs layer-wise decomposition of visual, textual, and historical attentions. Based on the VTACR distribution, Owl adaptively modulates attention along: a visual-favored path (enhancing grounding) and a text-favored path (amplifying hallucination). A dual-path contrastive decoding strategy then drives the LVLM to suppress hallucinations (e.g., Football) while preserving truthful predictions.

Mediator Analysis. Figure 1 (c) illustrates the results of causal interventions on modality-specific attention. Increasing visual attention weights A_V leads to a measurable reduction in hallucination scores $H(Y_T)$, confirming a positive TCE from A_V to the output Y_T. This indicates that enhancing A_V effectively strengthens the causal path $X_V \rightarrow A_V \rightarrow Y_T$, thereby reducing the influence of biased visual priors P_V and promoting image-grounded reasoning. In contrast, increasing textual attention A_T results in elevated $H(Y_T)$, suggesting that amplifying linguistic priors P_T can disrupt modality alignment and induce hallucinations. Conversely, suppressing A_T helps mitigate this effect. These findings validate A_V and A_T as mediators in the causal paths $P_V \rightarrow A_V \rightarrow Y_T$ and $P_T \rightarrow A_T \rightarrow Y_T$. Through soft interventions $do(A_V = A_V^*)$, $do(A_T = A_T^*)$, we demonstrate that attention allocation serves not only an architectural function but also a causal mechanism for controlling hallucinations in vision-language reasoning.

Adaptive Attention Modulation. To enable fine-grained hallucination suppression, we leverage the proposed VTACR to adaptively apply attention re-weighting. Specifically, we randomly sample $2,000$ hallucinated samples from MSCOCO (Lin et al. 2014) and compute the VTACR value $V^{(\ell)}$ via Equation (3) for hallucinated tokens in each decoder layer ℓ. By aggregating these values, we estimate a layer-wise VTACR density distribution. For each layer ℓ, we define the base score $V_b^{(\ell)}$ as the distribution's τ-th percentile, where τ is a hyperparameter. During generation, we obtain per-layer $V^{(\ell)}$ for the current token. If $V^{(\ell)} < V_b^{(\ell)}$, indicating insufficient visual grounding, we increase the visual/textual attention coefficient α, β. Otherwise, we retain their original values. This modulation is formulated as:

$$\tilde{T}^{(\ell)} = \mathbb{I}(V^{(\ell)} < V_b^{(\ell)}) \cdot \min\left(T \cdot \frac{V^{(\ell)} - V_b^{(\ell)}}{V_b^{(\ell)}}, T\right), \quad (9)$$

$$\tilde{\alpha}^{(\ell)} = \alpha + \tilde{T}^{(\ell)}, \quad (10)$$

$$\tilde{\beta}^{(\ell)} = \beta + \tilde{T}^{(\ell)}, \quad (11)$$

where T is a pre-defined modulation coefficient, and $\mathbb{I}(\cdot)$ is an indicator function that activates intervention only when
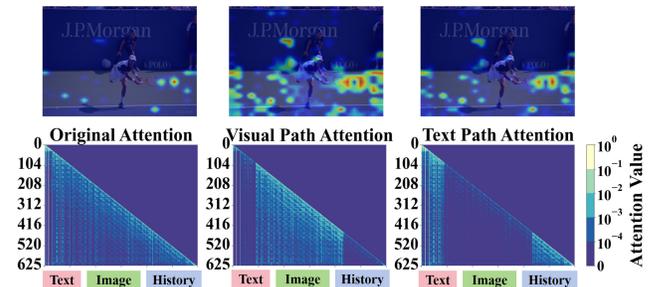

Figure 4: Visualization of dual-path attention intervention. Compared to original attention, the visual- and text-favored paths highlight distinct modality preferences in token-level.

the current $V^{(\ell)} < V_b^{(\ell)}$. This mechanism allows the model to dynamically prioritize visual/text features when hallucination risk is detected, while avoiding unnecessary intervention for well-grounded tokens.

Dual-path Attention Intervention and Contrastive Decoding. Unlike PAI that perform uniform intervention at the attention layer, our analysis based on the VTACR signal reveals a critical insight: the influence of visual and textual attention on hallucination varies significantly across different layers. Therefore, we propose attention intervention based on the contribution of each layer. We disentangle modality dominance by creating two hypothetical but informative paths: a *visual-favored* decoding path that amplifies visual grounding, and a *text-favored* path that mimics excessive textual reliance. This dual-path formulation enables us to assess and contrast the causal impact of A_V and A_T, and make more targeted corrections.

In the *visual-favored path*, we enhance the attention on visual tokens A_V and attenuate those on textual tokens A_T, forcing the model to rely more on image evidence:

$$\tilde{\mathbf{A}}_{i,j}^{(\ell)} = \mathbf{A}_{i,j}^{(\ell)} + \tilde{\alpha}^{(\ell)} \cdot |\mathbf{A}_{i,j}^{(\ell)}|, \quad (j \in \mathcal{V}) \quad (12)$$

$$\tilde{\mathbf{A}}_{i,k}^{(\ell)} = \mathbf{A}_{i,k}^{(\ell)} - \tilde{\beta}^{(\ell)} \cdot |\mathbf{A}_{i,k}^{(\ell)}|, \quad (k \in \mathcal{T}) \quad (13)$$

Conversely, the *text-favored path* downplays visual grounding and encourages decoding dominated by textual priors, simulating hallucination risks:

$$\tilde{\mathbf{A}}_{i,j}^{(\ell)} = \mathbf{A}_{i,j}^{(\ell)} - \tilde{\alpha}^{(\ell)} \cdot |\mathbf{A}_{i,j}^{(\ell)}|, \quad (j \in \mathcal{V}) \qquad (14)$$

$$\tilde{\mathbf{A}}_{i,k}^{(\ell)} = \mathbf{A}_{i,k}^{(\ell)} + \tilde{\beta}^{(\ell)} \cdot |\mathbf{A}_{i,k}^{(\ell)}|, \quad (k \in \mathcal{T}) \qquad (15)$$

where an example is visualized in Figure 4. The original attention exhibits dispersed focus over both text and image tokens, leading to ambiguous grounding. Under our VTACR-guided modulation, the visual-favored path sharpens attention toward salient image regions (e.g., the player and racket), while the text-favored path highlights prior textual tokens, potentially driving hallucinations.

Above two adaptive attention intervention paths produce logit distributions that reflect divergent modality dependencies. Based on this, we propose a Dual-path Contrastive Decoding (DCD) that explicitly simulates and contrasts two complementary decoding trajectories. By contrasting them, we can detect and suppress hallucination-prone outputs. The final prediction is obtained via contrastive fusion:

$$P_{\text{DCD}}(Y|X_V, X_I) = \text{Softmax}\Big[(1+\lambda) \cdot \log p_\theta(y|X_{V\uparrow}, X_{T\downarrow})$$
$$- \lambda \cdot \log p_\theta(y|X_{V\downarrow}, X_{T\uparrow})\Big] \qquad (16)$$

where $(X_{V\uparrow}, X_{T\downarrow})$ and $(X_{V\downarrow}, X_{T\uparrow})$ denote the visual and text favored decoding settings, respectively, and λ is a tunable contrastive strength. Combined with VTACR-guided attention calibration, this dual-path approach completes a causal intervention loop: measuring, adjusting, and decoding – all guided by mediator behavior.

5 Experiments

Benchmarks. We test Owl on two hallucination detection benchmarks and one comprehensive evaluation metric:

(1) POPE (Li et al. 2023c): The Polling-based Object Probing Evaluation assesses object hallucinations through a binary QA-style probing interface. It queries the model about the presence of specific objects in an image, bypassing the need for caption parsing, and providing a reliable, model-agnostic measurement.

(2) CHAIR (Rohrbach et al. 2018): The Caption Hallucination Assessment with Image Relevance evaluates hallucinations at both the instance and sentence levels. CHAIR$_I$ quantifies the proportion of hallucinated objects among all mentioned objects, while CHAIR$_S$ measures the percentage of captions containing at least one hallucinated object.

(3) GPT-4V Assisted Evaluation: Following (Huang et al. 2024b; Liu, Zheng, and Chen 2024), we adopt GPT-4V (Yang et al. 2023) to comprehensively evaluate the semantic correctness and visual faithfulness of the generated captions.

Models and Baselines. We evaluate our approach on three representative LVLM backbones, each covering a distinct architectural paradigm: **LLaVA-1.5** (Liu et al. 2024a): An alignment-optimized vision-language model based on

CLIP and Vicuna; **MiniGPT-4** (Zhu et al. 2023): A task-agnostic LVLM with Vicuna-based decoding and BLIP-2-style alignment; **Shikra** (Chen et al. 2023): A structured LVLM that supports object-level localization and fine-grained grounding. For comparison, we include both classic decoding strategies – Beam Search, Greedy Decoding, and Nucleus Sampling – and several state-of-the-art baselines: **VCD** (Leng et al. 2024): Introduces visual contrastive decoding to suppress language bias and enhance visual grounding. **PAI** (Liu, Zheng, and Chen 2024): Modulates attention via perplexity-aware gating to improve robustness under ambiguous conditions. **OPERA** (Huang et al. 2024b): Prevents repetitive hallucinations through rollback and attention weight suppression. **CausalMM** (Zhou et al. 2025): Utilizes a causal diagram to apply counterfactual reasoning to both the vision encoder and language decoder.

Configurations and Parameters. All experiments are conducted on 500 images randomly sampled from the MSCOCO val2014 dataset (Lin et al. 2014), following the setup in prior works (Huang et al. 2024b; Liu, Zheng, and Chen 2024). The visual/textual attention coefficient α, β are empirically tuned for each model to balance the quality of the generation and the reduction of hallucinations. Specifically, we set $(\alpha=0.4, \beta=0.5)$ for LLaVA-1.5, $(0.2, 0.3)$ for MiniGPT-4, and $(0.5, 0.3)$ for Shikra. The contrastive decoding strength λ is fixed at 0.2, the modulation coefficient T is set to 0.2 across all models, and the default value of τ is 80. All experiments are conducted on $4\times$NVIDIA 3090 GPUs. The reported results are the best of 20 runs for all models and the statistical significance of the results is less than 0.05, i.e., $p < 0.05$.

Results on CHAIR hallucination evaluation.

Table 1 shows Owl consistently outperforms all baselines across three LVLMs in both sentence-level (C_S) and instance-level (C_I) hallucination metrics. Compared to the strongest prior method PAI, our Owl achieves substantial C_S reductions of 17.6%, 14.5%, and 22.1%, and C_I reductions of 21.4%, 36.7%, and 24.8% on LLaVA-1.5, MiniGPT-4, and Shikra, respectively. Importantly, this hallucination suppression is achieved while preserving or even improving generation length, indicating that our intervention does not compromise output richness. These gains stem from our adaptive, VTACR-guided dual-path decoding strategy. Unlike PAI or OPERA that apply static or uni-modal atten-

Method	LLAVA-1.5			MiniGPT-4			Shikra		
	$C_S \downarrow$	$C_I \downarrow$	Len↑	$C_S \downarrow$	$C_I \downarrow$	Len↑	$C_S \downarrow$	$C_I \downarrow$	Len↑
Beam Search	48.6	16.2	105.3	33.0	8.7	77.8	53.8	17.4	116.2
Greedy	47.4	16.4	108.7	34.2	8.8	79.5	51.9	16.9	118.9
Nucleus	47.3	16.5	107.1	32.1	7.6	78.3	54.5	17.8	117.5
VCD	44.6	14.4	93.8	33.1	11.2	71.2	47.2	15.5	105.6
OPERA	42.2	13.1	89.5	30.1	9.8	68.7	36.8	12.4	98.3
PAI	31.8	10.3	85.2	24.8	9.3	65.9	37.6	12.9	94.7
CausalMM	35.8	12.3	87.6	27.8	9.6	67.3	41.6	13.4	96.8
Ours	**26.2**	**8.1**	**98.4**	**21.2**	**6.2**	**73.6**	**29.3**	**9.7**	**108.2**

Table 1: Results on CHAIR. C_S and C_I denote CHAIR$_S$ and CHAIR$_I$. Len is the average length of generated text.

Method	LLAVA-1.5			MiniGPT-4			Shikra		
	Ran↑	Pop↑	Adv↑	Ran↑	Pop↑	Adv↑	Ran↑	Pop↑	Adv↑
Beam Search	84.6	84.4	83.1	69.2	68.8	67.4	81.5	78.1	79.2
Greedy	83.6	84.2	80.2	64.6	63.3	62.2	81.9	78.1	80.2
Nucleus	79.4	78.2	76.6	62.8	59.8	58.5	80.2	76.5	78.2
VCD	86.2	87.1	87.5	73.9	71.2	70.3	81.2	77.3	80.8
OPERA	87.6	88.2	90.1	75.8	74.8	72.4	83.5	79.2	82.1
PAI	89.8	**89.3**	90.3	78.3	**79.8**	78.9	81.5	79.2	80.5
CausalMM	88.1	87.6	82.9	74.1	73.6	75.2	83.4	79.6	81.7
Owl(Ours)	**90.2**	88.1	**90.5**	**82.2**	78.4	**79.0**	**85.2**	**82.3**	**83.4**

Table 2: Results on POPE. Ran, Pop, and Adv is an abbreviation for *Random*, *Popular*, and *Adversarial* setting, respectively. The higher score indicates better performance.

Figure 6: Impact of α, β, and λ on hallucination and informativeness in LLaVA-1.5, evaluated on 500 COCO samples.
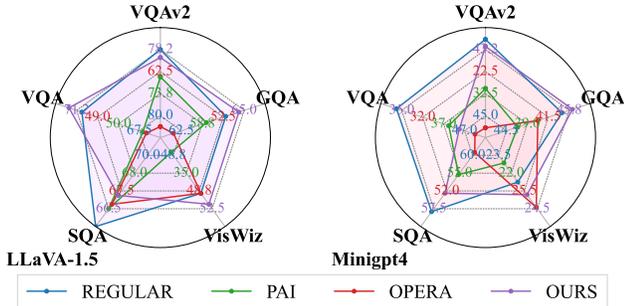

Figure 5: Comparison among different VLMs on five VQA benchmarks and three common benchmarks. The highest-performing results are highlighted in boldface.

tion shifts, our approach dynamically adjusts visual and textual attention at each decoding layer based on token-level VTACR scores. This enables effective hallucination suppression without overcorrecting or truncating informative content – striking a better balance between grounding and fluency. Notably, though CausalMM intervenes attention in both visual encoder and LLM layers, it inherently amplifies hallucinatory signals instead of enhancing image-awareness. Unlike our DCD, which sufficiently widens the gap between faithful and hallucination tokens, this limits its efficacy.

Results on POPE generalization benchmark. Table 2 reports results under Random, Popular, and Adversarial settings. Owl consistently outperforms classic decoding baselines (Beam Search, Greedy, Nucleus) across all LVLMs and splits, with particularly strong gains in adversarial scenarios. Besides, Owl outperforms other hallucination mitigation methods in a competitive manner in most cases. While slightly trailing PAI on MiniGPT-4 and LLaVA-1.5 under the Popular setting, we speculate that this setting prioritizes high-frequency objects, which aligns better with PAI's scenario of textual inertia. For Shikra, Owl delivers the highest accuracy in three settings. These improvements highlight the robustness of Owl under varied linguistic priors. By explicitly contrasting visually grounded and hallucination-prone paths, Owl enhances model generalization.

Results on vision-language understanding ability. To assess whether our hallucination mitigation hurts general vision-language understanding, we evaluate performance on five VQA benchmarks: VQAv2 (Goyal et al. 2017), GQA (Hudson and Manning 2019), VizWiz (Gu-
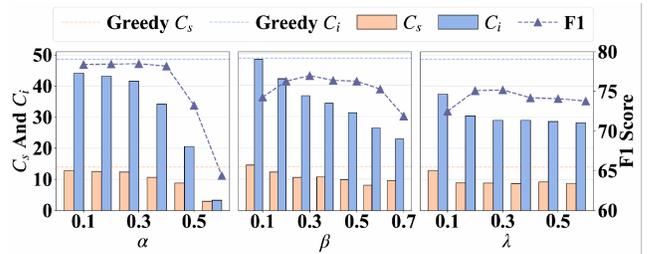
rari et al. 2018), ScienceQA-IMG (Lu et al. 2022), and TextVQA (Singh et al. 2019). As illustrated in Figure 5, Owl achieves consistent performance with the base LVLMs and even outperforms them on several benchmarks. For example, on LLaVA-1.5 compared to regular, we observe TextVQA (+3.7). Notably, gains on VizWiz (from 48.8 to 52.5, 7.6% ↑), where understanding visually degraded or text-heavy content is crucial. This suggests that our visual-attention-enhancing interventions help the model better localize and utilize visual cues under challenging conditions. Meanwhile, only a marginal decrease (from 80.0 to 78.2, 2.3% ↓) is observed on VQAv2, indicating minimal trade-offs in general capability. Similar trends hold on MiniGPT-4, with a notable gain on GQA from 44.5 to 45.8 (+1.3), and a slight decrease on VQA-v2 from 45.0 to 43.2 (−1.8), validating that our approach preserves or enhances reasoning under fine-grained or noisy visual contexts. Overall, these results confirm that our hallucination mitigation strategy maintains core vision-language understanding and even benefits tasks requiring robust visual grounding, thanks to our causal attention modulation and dual-path decoding.

Ablation Study. We conduct ablations to assess the impact of three key hyperparameters in our DCD strategy: visual/textual attention coefficients α, β, and the contrastive decoding strength λ. Evaluation is based on CHAIR for hallucination and F1 score for information richness and accuracy (Liu, Zheng, and Chen 2024). As shown in the LLaVA-1.5 results in Figure 6, increasing α in Equations (12) and (14) enhances visual grounding and reduces hallucinations, but overly large values suppress informative content, lowering F1 score – revealing a trade-off between hallucination mitigation and content richness. Increasing β in Equations (13) and (15) steadily reduces hallucinations with a minimal drop in F1 score, suggesting that stricter regulation of textual attention effectively counters language priors without harming visual relevance, since the DCD strategy widens the gap between hallucination tokens and faithful tokens from Equation (16). The parameter λ balances visual and textual interventions during decoding. Moderate values (0.1–0.4) yield stable improvements, while excessively high λ harms both CHAIR and F1, indicating decoding instability. In general, the results demonstrate that α, β, and λ play distinct but complementary roles. Careful tuning enables a balanced trade-off between hallucination suppression and semantic completeness.
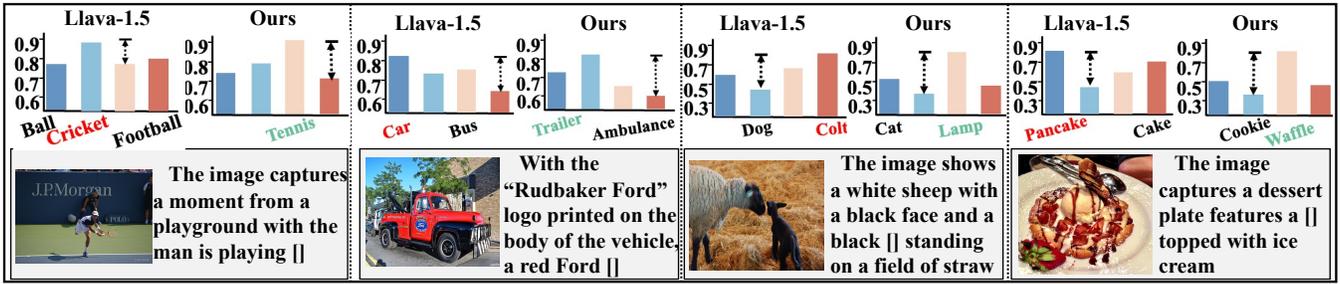
Figure 7: Visualization of token logits at hallucination-prone positions []. Red bars is hallucinated tokens and greens denote faithful ones. Our Owl consistently suppresses hallucinations and promotes visually-grounded predictions via DCD strategy.
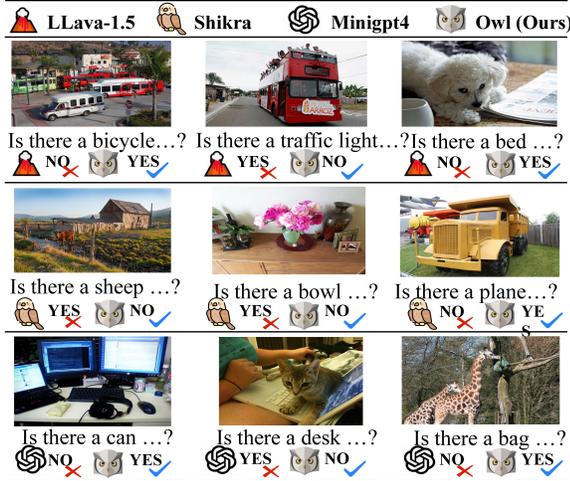

Figure 8: Comparison on the POPE benchmark. Our Owl reliably avoids hallucinations and yields accurate predictions across diverse models and scenes.
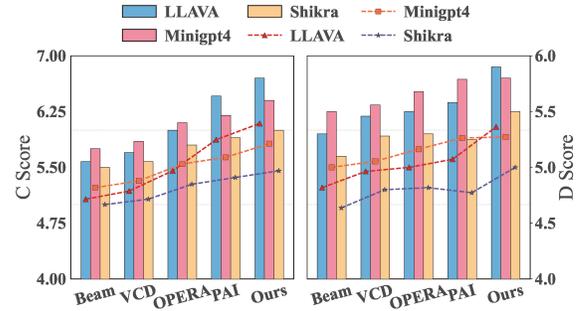

Figure 9: GPT-4V hallucination evaluation on MSCOCO. Left: Correctness (higher = less hallucination); Right: Detailedness. Line plots show model score trends per method.

Case study. To further illustrate the effectiveness of Owl, we present two qualitative analyses. In Figure 7, we visualize the Top-4 token logits predicted by LLaVA-1.5 with and without our Owl, which demonstrates how Owl consistently suppresses hallucinated tokens and promotes visually grounded predictions by adjusting attention distribution during decoding. Compared to LLaVA-1.5, which often assigns higher logits to misleading tokens influenced by language priors, our approach reweights token importance through visual enhancement and textual suppression, yielding more accurate results. In Figure 8, we visualize examples from the POPE benchmark, comparing Owl with three backbones. Although existing models frequently misidentify nonexistent objects (e.g. "traffic light", "bowl," or "bag"), Owl produces consistently correct responses. This highlights the strength of Owl framework in resisting hallucinations, particularly in cases where language priors strongly conflict with visual evidence.

Results on human-like GPT-4v assisted hallucination evaluation. Figure 9 presents GPT-4V evaluations of Owl against beam search, VCD, OPERA, and PAI across three backbone models. On LLaVA-1.5, our method lifts Correctness from 5.58 to 6.70 – a notable improvement of

20.1% – and enhances Detailedness from 5.30 to 5.90 (up by 11.3%). For MiniGPT-4, Correctness increases from 5.75 to 6.40 (11.3% ↑), and Detailedness grows from 5.50 to 5.80 (5.5% ↑). Shikra shows a Correctness gain from 5.50 to 6.00 (9.1% ↑), with Detailedness rising from 5.10 to 5.50 (7.8% ↑). Across all backbones, Owl consistently improves Correctness while maintaining or slightly enhancing Detailedness, with LLaVA-1.5 exhibiting the most significant gains. These results confirm that our Owl framework effectively reduces hallucinations without compromising the richness of the generated content.

6 Conclusion

In this work, we present a causally grounded framework **Owl** for mitigating object hallucinations in LVLMs. By modeling the generation process through a structural causal model with visual and textual attention as mediators, we uncover the causal roots of hallucination and propose VTACR – a novel metric to quantify the cross-modal attention balance during decoding. Guided by VTACR signals, we develop fine-grained attention interventions and a dual-path contrastive decoding strategy that effectively suppress hallucinated content. Extensive experiments on POPE and CHAIR benchmarks validate our approach, achieving a 22.9% reduction in hallucination rates over strong baselines.

This work provides both theoretical insights and practical tools for enhancing LVLM faithfulness, and opens new avenues for causal control in multimodal generation.

Acknowledgements

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, J.; Guo, H.; Peng, Z.; Yang, J.; Li, Z.; Li, M.; and Tian, Z. 2025. Mitigating hallucinations in large vision-language models by adaptively constraining information flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 23442–23450.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Deng, A.; Chen, Z.; and Hooi, B. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18135–18143.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.

Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; and Luo, P. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22170–22183.

Huang, P.-H.; Li, J.-L.; Chen, C.-P.; Chang, M.-C.; and Chen, W.-C. 2024a. Who Brings the Frisbee: Probing Hidden Hallucination Factors in Large Vision-Language Model via Causality Analysis. *arXiv preprint arXiv:2412.02946*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024b. 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*, 13418–13427.

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

Jiang, Z.; Chen, J.; Zhu, B.; Luo, T.; Shen, Y.; and Yang, X. 2024. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41.

Kim, B. S.; Kim, J.; Lee, D.; and Jang, B. 2025. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Computing Surveys*, 57(10): 1–35.

Lange, B.; Yildiz, A.; Arief, M.; Khattak, S.; Kochenderfer, M.; and Georgakis, G. 2025. General-Purpose Robotic Navigation via LVLM-Orchestrated Perception, Reasoning, and Acting. *arXiv preprint arXiv:2506.17462*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, W.; Su, X.; Song, D.; Wang, L.; Zhang, K.; and Liu, A.-A. 2023b. Towards deconfounded image-text matching with causal inference. In *Proceedings of the 31st ACM international conference on multimedia*, 6264–6273.

Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; and Lewis, M. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *URL https://arxiv. org/abs/2407.21771*.

Liu, Z.; Chu, T.; Zang, Y.; Wei, X.; Dong, X.; Zhang, P.; Liang, Z.; Xiong, Y.; Qiao, Y.; Lin, D.; et al. 2024b. Mmdu: A multi-turn multi-image dialog understanding benchmark

and instruction-tuning dataset for lvlms. *Advances in Neural Information Processing Systems*, 37: 8698–8733.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521.

Lyu, X.; Chen, B.; Gao, L.; Song, J.; and Shen, H. T. 2024. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *arXiv preprint arXiv:2405.15356*.

Lyu, Y.; Yang, Z.; Niu, Y.; Jiang, J.; and Lo, D. 2025. Do Existing Testing Tools Really Uncover Gender Bias in Text-to-Image Models? In *Proceedings of the 33rd ACM International Conference on Multimedia*.

Neal, B. 2020. Introduction to causal inference. *Course lecture notes (draft)*, 132.

Pearl, J. 2010. Causal inference. *Causality: objectives and assessment*, 39–58.

Pearl, J. 2022. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, 373–392. Cambridge University Press.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12): 220105.

Yu, H.; Qiu, Y.; Yang, Y.; Fang, H.; Zhuang, T.; Hong, J.; Chen, B.; Wu, H.; and Xia, S.-T. 2025a. ICAS: Detecting Training Data from Autoregressive Image Generative Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 11209–11217.

Yu, L.; Guo, L.; Kuang, P.; and Zhou, F. 2025b. Bridging the fairness gap: Enhancing pre-trained models with llm-generated sentences. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Yu, L.; Sun, J.; Kuang, P.; Zhou, R.; Zhou, F.; and Feng, Z. 2025c. Bimodal Debiasing for Text-to-Image Diffusion: Adaptive Guidance in Textual and Visual Spaces. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 11249–11258.

Zhang, C.; Zhang, L.; Wu, J.; He, Y.; and Zhou, D. 2025a. Causal prompting: Debiasing large language model prompting based on front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25842–25850.

Zhang, F.; Zhang, Z.; Zhang, X.; and Xu, C. 2025b. When Open-Vocabulary Visual Question Answering Meets Causal Adapter: Benchmark and Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9887–9895.

Zhang, X.; Quan, Y.; Gu, C.; Shen, C.; Yuan, X.; Yan, S.; Cheng, H.; Wu, K.; and Ye, J. 2024. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*.

Zhou, F.; Mao, Y.; Yu, L.; Yang, Y.; and Zhong, T. 2023a. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4227–4241.

Zhou, G.; Yan, Y.; Zou, X.; Wang, K.; Liu, A.; and Hu, X. 2025. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *ICLR*.

Zhou, J.; Gao, J.; Zhao, X.; Yao, X.; and Wei, X. 2024. Association of objects may engender stereotypes: Mitigating association-engendered stereotypes in text-to-image generation. *Advances in Neural Information Processing Systems*, 37: 51754–51786.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2023b. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.