TokenPenalty: Alleviating Attention Sinks and Positional Decay in LVLMs

Xiaofeng Zhang*[1], Yuanchao Zhu*[1], Qiyan Zhao[1], Xiaosong Yuan[2]†, JiaWei Cao[1], Xuhang Chen[3]†
[1]Department of Automation and Intelligent Sensing, Shanghai Jiao Tong University
[2]Jilin University [3]Huizhou University
{framebreak@}sjtu.edu.cn

Abstract

Large vision language models (LVLMs) are increasingly deployed in Web-scale applications—such as image search, social media captioning, and e-commerce product description generation—where factual consistency is critical for user trust and content reliability. However, we observe that LVLMs frequently hallucinate in these settings due to two relevant phenomena: the massive activation phenomenon and positional information decay. The former refers to the tendency of attention mechanisms to concentrate on a small set of tokens with extreme activation values in query and key projections, which play indispensable roles in contextual understanding. In our investigation, we discover that perturbing these tokens leads to significant performance drops, highlighting their utmost importance. As for positional information decay, it occurs due to the common rotary position encoding strategy, where the attention to early visual tokens diminishes over time, especially in long-sequence generation tasks, such as image caption. To address these challenges, we propose **TokenTruth**, a token-level intervention strategy that dynamically suppresses irrelevant visual tokens while preserving key contextual signals. Our method is grounded in an in-depth analysis of massive activations and attention sink behaviors, and introduces a targeted token penalty mechanism that reallocates attention more faithfully toward informative visual regions. Extensive experiments demonstrate that TokenTruth significantly improves factual consistency across various LVLMs on standard image understanding benchmarks.

1 Introduction

Large vision language models(LVLMs) (Achiam et al., 2023; Liu et al., 2024a; Bai et al., 2023; Bi et al., 2025a,b; Wang et al., 2025b; Jiang et al., 2025b,a; Rong et al., 2025; Zhang et al., 2023; Peng et al., 2025) have achieved remarkable progress in various cross-modal tasks, particularly in the integration of text and image modalities. Despite these advances, hallucinations within LVLMs remain a significant challenge, especially in tasks such as visual question answering (VQA) and image captioning.

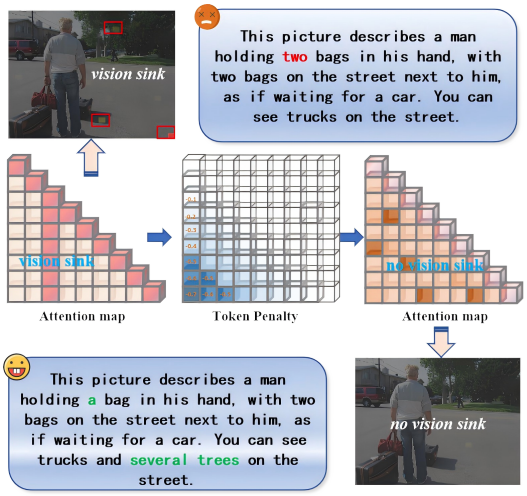

Figure 1: The deep vision attention in the model is dynamically allocated, alleviating the attention sink after the TokenTruth intervention,

Recent studies on attention sinks have provided novel perspectives on understanding hallucinations. The concept of attention sink as an information flow is introduced in Label Words are Anchors (Wang et al., 2023b), which demonstrates how information flows in large language models (LLMs) frequently converges on specific user tokens. Based on this, OPERA (Huang et al., 2024), DOPRA (Wei and Zhang, 2024), TAME (Tang et al., 2025a), middle (Jiang et al., 2024b), EAH (Zhang et al., 2024a),Farsight (Tang et al., 2025b) and others (?Zhang et al., 2026a; Zhao et al., 2026; Zhang et al., 2026b) investigate the relationship between

† Corresponding author
* These authors contributed equally to this work

Findings of the Association for Computational Linguistics: ACL 2026, pages 12894–12906
July 2-7, 2026

the attention sink in different tokens in LVLM. They discover that when a token consistently attracts high attention weights across subsequent tokens, such an over-reliance can lead to hallucinations in the model's output. Although these studies have explored the attention sink phenomenon for various text tokens, the relationship between the attention sink, image tokens, and hallucinations is still unclear.

***Background**: Massive activation tokens are crucial for contextual understanding.*

In the continued exploration of attention sink work, Jin (Jin et al., 2025) found that Rotary Position Embedding (RoPE) induces a strong concentration of maximum activation values in the query (Q) and key (K) projections, with no such pattern observed in the value (V) representations. These extreme values in Q and K are essential for contextual understanding, primarily influencing the model's ability to process information within the current context window rather than relying on stored knowledge from parameters. On tasks demanding contextual coherence, perturbing these maxima results in severe performance drops, underscoring their functional importance in attention-based architectures.

***Question**: How does the attention-sink caused by the maximum activation value of the image token affect contextual understanding tasks such as image caption?*

As shown in Figure 2, to verify the role of the maximum activation value in LVLM's context understanding, we investigate the method of directly intervening in the attention of the sink token, such as vissink (Kang et al., 2025) that intervenes in the token corresponding to the maximum activation value. We notice that the vissink method can effectively improve models on some QA datasets, such as GQA and SQA, while reducing the performance for long output datasets with contextual understanding and output such as CHAIR. This can also be validated from another aspect, where the anchor token corresponding to the attention sink plays a key role in contextual understanding. Therefore, reallocating attention to such anchor tokens directly will damage the performance on image caption, etc.

This suggests that a more nuanced and dynamic

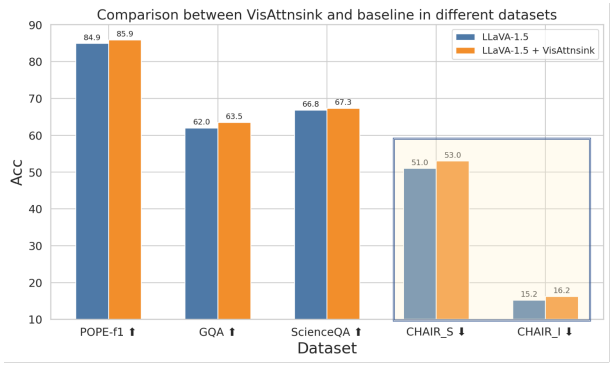

Figure 2: Comparing the results of Vissink (Kang et al., 2025) with the baseline, we can find that it is effective in short-answer QA questions such as GQA and SQA, but the effect of long-answer description tasks such as CHAIR is reduced, which shows that directly removing or averaging the vision attention sink will cause contextual information loss.

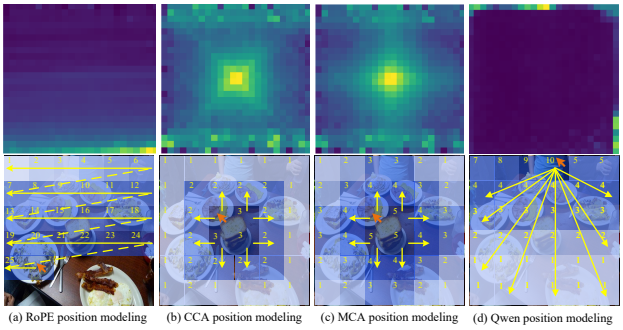

Figure 3: Long-distance position information decay of qwen2-vl and llava, (a) represents the long-distance information attenuation problem in LLaVA, (b) (c) represent CCA-LLaVA (Xing et al., 2024a) and MCA-LLaVA (Zhao et al., 2025) to alleviate the decay of long-distance information flow by optimizing RoPE, (d) represents the position decay of Qwen2-VL.

approach is needed to handle the distribution of attention across vision tokens, rather than simply focusing on predefined anchor tokens. Notably, we observe a progressive attenuation of visual information in Qwen2-VL and LLaVA1.5 (As shown in Figure 3) .during long-sequence generation. We refer to this as positional information decay, which has also been mentioned in CCA-LLaVA (Xing et al., 2024a), Farsight (Tang et al., 2025b), and MCA-LLaVA (Zhao et al., 2025). This occurs because the attention mechanism tends to favor more recent tokens, causing earlier visual tokens to lose influence over time.

To this end, we introduce **TokenTruth**, which consists of a token-level penalty and a head-level enhancement. Token-level penalty adapts to the

evolving context during decoding, ensuring that critical information from both visual and textual modalities is preserved and propagated effectively. We designed the AliBi bias penalty to not only alleviate the sinking of visual attention, but also alleviate the long-distance decay phenomenon. By incorporating this dynamic attention allocation strategy, we effectively absorb and redistribute excessive attention from low-informative or misleading tokens, ensuring that key contextual signals are preserved and propagated across the sequence. This strategy is particularly beneficial in long-sequence generation tasks, where the challenge of maintaining coherence while retaining input information escalates with increasing sequence length.

With extensive experiments, our method demonstrates significant performance across several LVLMs on image hallucination and generation benchmarks, proving its effectiveness. Specifically, our contributions can be summarized as follows:

- This paper reveals and demonstrates that the massive value corresponding to the vision attention sink token is crucial for understanding contextual knowledge, and direct intervention will destroy the ability of tasks such as the image caption task.
- This paper proposes a two-stage attention allocation strategy named **TokenTruth**, the first stage intervenes in the propagation process between token levels by optimizing the causal mask, and the second stage intervenes in the head-level attention to redistribute, thereby enhancing the contextual reasoning ability.
- Through extensive experiments on multiple benchmarks, we demonstrate the consistent effectiveness of TokenTruth in improving visual grounding and reducing hallucination.

2 Related Work

2.1 Attention Sink and Information Flow

With the rapid development of LVLMs, extensive research has been conducted to find inspiration for model optimization by analyzing the internal mechanisms of the models. Among them, the information flow (Wang et al., 2023b) provides an intuitive method to understand the internal mechanisms of the LVLM black-box models. Label Words (Wang et al., 2023b) and ACT (Yu et al., 2024) are early works that explore the mechanism of LLMs(Zhu et al., 2023; Devlin et al., 2018; Touvron et al., 2023) by observing information flow patterns. By calculating saliency scores, it is possible to visualize the information flow.

The label words introduce the information flow (Wang et al., 2023b). The authors define the saliency score as the product of the attention score and the gradient. Tokens with significant aggregation in saliency scores are identified as anchor tokens.

OPERA (Huang et al., 2024) and DOPRA (Wei and Zhang, 2024) identify that anchor output tokens can lead to hallucinated token generation and try to penalize anchor tokens' logits. Furthermore, FastV (Chen et al., 2024b) addresses inefficiencies in attention calculations, particularly in deeper layers of models like LLaVA-1.5, employing an image token pruning strategy to accelerate inference without compromising performance. Zhang (Zhang et al., 2024b,d, 2025b) uses Grad-CAM and attention maps to visualize the interaction between images and text in complex reasoning tasks. The EAH (Zhang et al., 2024a) indicates that most hallucinations stem from the attention sink pattern marked by images in the attention matrix. TAME (Tang et al., 2025a) investigates the causes of hallucinations by analyzing local self-attention patterns of "anchor points" and defines the degree of attentional localization as the probability of marker propagation. Massive (Sun et al., 2024) highlights that while there are approximately 40,000 activations per hidden state, only four are recognized.

Inspired by causal reasoning in decoding strategies, Farsight (Tang et al., 2025b) proposes to use causal masks to establish information propagation between multimodal words.

2.2 Difference between TokenTruth and Concurrent Work

Our approach differs from concurrent methods in several key aspects:

(1)vs. Farsight: While Farsight uses static causal masks, our Token Penalty provides dynamic, context-aware redistribution

(2) vs. EAH: EAH replaces attention heads entirely, while our Head Enhancement preserves original semantics through controlled redistribution.

(3) vs. Vissink: Vissink redistributes attention on BOS tokens, while we systematically handle attention patterns at the image token level and attention head level.

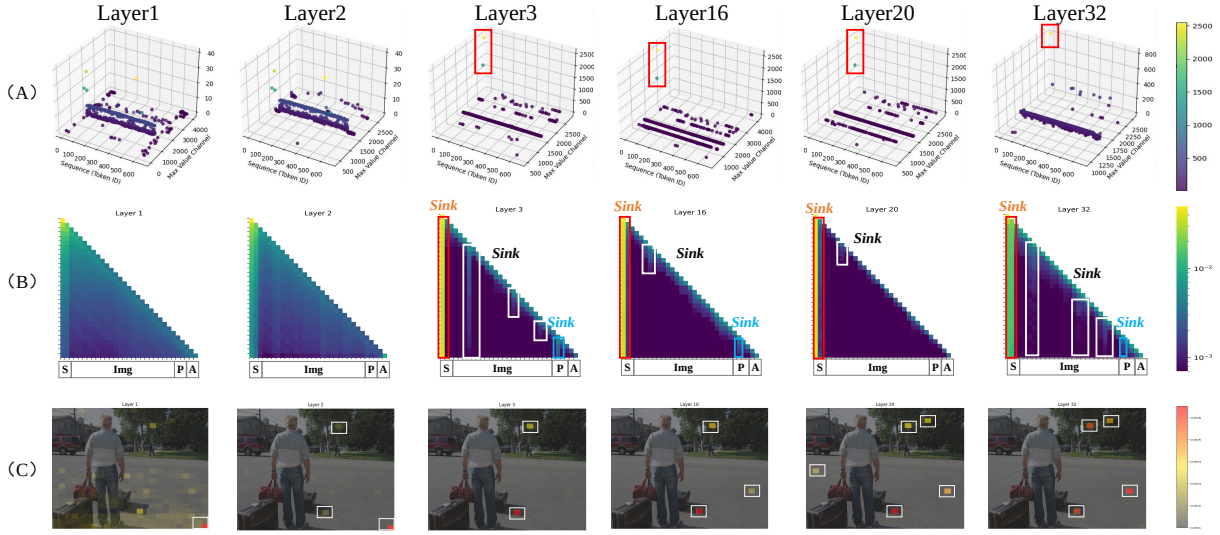

Figure 4: Progressive attention sink formation and massive activation patterns across different layers. (a) Channel-wise activation magnitude distribution, (b) Attention weight matrices showing sink formation, (c) Spatial attention visualization highlighting problematic regions, baseline: LLaVA1.5-7B.

2.3 Strategies for Solving the LVLM Hallucination Problem

The term "hallucination" (Zhang et al., 2024c, 2025c; Zhao et al., 2026; Zhang et al., 2026b; Gao et al., 2025; Sarkar et al., 2024; Xiao et al., 2024; Xing et al., 2024b; Ma et al., 2024; Gong et al., 2024a; Chen et al., 2024a; Kim et al., 2024; Liu et al., 2024b; Zhou et al., 2023; Zhai et al., 2023; Wang et al., 2023a; Huang et al., 2023; Zhu et al., 2024; Jiang et al., 2024a; Zhou et al., 2025; Bai et al., 2025; Suo et al., 2025; Lymperaiou et al., 2025; Wang et al., 2025a; Li et al., 2025c; Chen et al., 2025b; Che et al., 2025; Chen et al., 2025a; Tu et al., 2025; Mao et al., 2025; Duan et al., 2025; Yin et al., 2025; Li et al., 2025d; Bae et al., 2025; Li et al., 2025a; Xie et al., 2025; Zheng et al., 2024; Kan et al., 2024; Zhao et al., 2024; Wang et al., 2024a,c; He et al., 2024; Yang et al., 2024; Gu et al., 2024; Neo and Chen, 2024; Li et al., 2025b,e; Zhang et al., 2025a; Wu et al., 2025a; Liu et al., 2024c; Zou et al., 2024; Gong et al., 2024b; Zhou et al., 2024; Shang et al., 2024) refers to the fact that multimodal models, when processing multimodal inputs, sometimes produce content that does not correspond to the actual inputs or is even fictitious. Among them, RLHF is an approach that relies on human feedback reinforcement learning techniques, which manually evaluates and guides model outputs, prompting the model to pay more attention to factual basis and logical consistency.

3 Analysis and Motivation

Attention Entropy and Hallucination Correlation. Let $\boldsymbol{\alpha}_i^{\ell,h} \in \mathbb{R}^n$ denote the attention weights for token i at layer ℓ and head h, where n is the sequence length. We define the attention distribution entropy as:

$$\Phi(\boldsymbol{\alpha}_i^{\ell,h}) = -\sum_{j=1}^{n} \alpha_{i,j}^{\ell,h} \log \alpha_{i,j}^{\ell,h}. \tag{1}$$

Based on empirical analysis in multiple LVLMs, we establish that the hallucination probability P_h for a given generation step exhibits an inverse relationship with attention entropy.

$$P_h \propto \exp(-\beta \Phi(\boldsymbol{\alpha}_i^{\ell,h})), \tag{2}$$

where $\beta > 0$ is a model-dependent scaling parameter that characterizes the sensitivity of hallucination to attention concentration.

Massive Activation Theorem. Building upon the findings of (Jin et al., 2025), we prove that tokens with extreme activation values in query-key projections are essential for contextual understanding. Formally, let $\mathcal{M}^{\ell,h} = \{j : |q_j^{\ell,h} \cdot k_j^{\ell,h}| > \tau_m\}$ denote the set of massive activation tokens at layer ℓ and head h, where τ_m is a threshold.

Massive Activation Preservation. For any context-dependent task, removing or significantly perturbing tokens in $\mathcal{M}^{\ell,h}$ results in performance degradation $\Delta P \geq \epsilon |\mathcal{M}^{\ell,h}|$, where $\epsilon > 0$ is task-dependent.

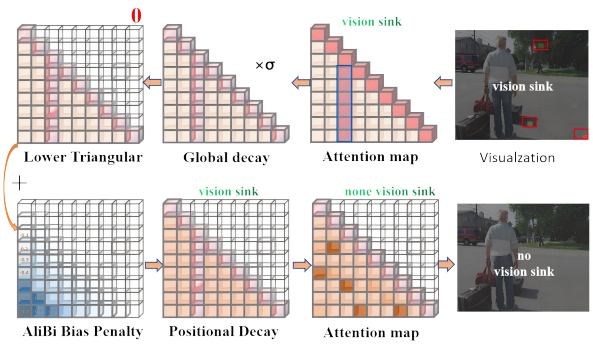

Figure 5: The structure of Token Level Penalty.

This theorem explains why direct intervention methods like VisSink (Kang et al., 2025) succeed in short-answer tasks but fail in contextual understanding tasks such as image captioning.

Positional Information Decay Model. For RoPE-based position encoding, we model the attention decay to early visual tokens as:

$$\alpha_{i,j}^{\ell,h} = \alpha_{i,j}^{\ell,h}(0) \cdot \exp(-\gamma \cdot |pos_i - pos_j|), \quad (3)$$

where pos_i, pos_j are positional indices, $\gamma > 0$ is the decay rate, and $\alpha_{i,j}^{\ell,h}(0)$ is the initial attention weight without positional bias.

Optimization Objective. Our two-stage approach aims to solve the constrained optimization problem:

$$\max_{\tilde{\boldsymbol{\alpha}}} \quad \Phi(\tilde{\boldsymbol{\alpha}}_i^{\ell,h}) \quad (4)$$

$$\text{s.t.} \quad \|\tilde{\boldsymbol{\alpha}}_i^{\ell,h} - \boldsymbol{\alpha}_i^{\ell,h}\|_{\mathcal{M}^{\ell,h}} \leq \delta \quad (5)$$

$$\sum_j \tilde{\alpha}_{i,j}^{\ell,h} = 1, \quad \tilde{\alpha}_{i,j}^{\ell,h} \geq 0 \quad (6)$$

where $\tilde{\boldsymbol{\alpha}}_i^{\ell,h}$ is the optimized attention distribution, $\|\cdot\|_{\mathcal{M}^{\ell,h}}$ denotes the norm restricted to massive activation tokens, and δ is a preservation constraint. The first constraint ensures that massive activation tokens retain their contextual importance, while the optimization objective increases attention entropy to reduce hallucination probability according to Equation 2.

4 Methodology

Building on the above analysis, we propose TokenTruth, a two-stage attention optimization framework (Figure 5): Token-Level Penalty mitigates attention collapse while preserving critical massive activations.

Listing 1 Implementation of the `Token Penalty` function for attention computation.

```
def TokenPenalty(scores, Key, Value, seq_len,
↪  sigma):
    # Step 1: Extract dimensions and build causal
    ↪  mask
    K = key_states.size(-2)
    C_full = torch.tril(torch.ones((K, K),
    ↪  device=scores.device, dtype=scores.dtype),
    ↪  diagonal=0)
    C_mask =
    ↪  C_full[-seq_len:].unsqueeze(0).unsqueeze(0)
    scores = scores * sigma * C_mask

    # Step 2: Build token level penalty Alibi bias
    idx_i = torch.arange(K - seq_len, K,
    ↪  device=scores.device).unsqueeze(1)
    idx_j = torch.arange(K,
    ↪  device=scores.device).unsqueeze(0)
    distance = (idx_j -
    ↪  idx_i).clamp(min=0).to(scores.dtype)

    pool_score = -distance.unsqueeze(0) *
    ↪  self.alibi_slopes
    scores = scores + pool_score.unsqueeze(0)

    # Step 3: Softmax normalization with causal
    ↪  masking
    attn_probs = torch.softmax(scores, dim=-1)
    attn_probs = attn_probs * C_mask

    # Step 4: Compute final output using values
    attn_output = torch.matmul(attn_probs,
    ↪  value_states)

    # Step 5: Final projection
    bsz, num_heads, _, head_dim = attn_output.size()
    attn_output = attn_output.transpose(1,
    ↪  2).reshape(bsz, seq_len, num_heads *
    ↪  head_dim)
    attn_output = self.o_proj(attn_output)

    return attn_output
```

4.1 Token Level Penalty

As shown in Figure 4, LVLMs like LLaVA and Qwen2-VL exhibit *attention collapse*—over-concentration on visual tokens due to RoPE-induced extreme activations (Jin et al., 2025). This harms visual grounding by violating the entropy maximization objective (Eq. 6). To address both issues, we introduce a token-level penalty mechanism based on the ALiBi biasing strategy, which dynamically penalizes distant and over-attended tokens. Let the raw attention score matrix be:

$$A_{i,j}^{\prime\ell,h} = \sigma\left(\frac{Q\,K^\top}{\sqrt{D_k}}\right), \quad (7)$$

Global Decay Factor: We introduce a learnable parameter $\sigma \in [0, 1]$ to scale the raw attention scores before applying the positional bias. This factor is optimized through ablation studies (Section 5.4) and set to $\sigma = 0.9$ across all experiments, balancing content relevance and positional information. To enforce causal (autoregressive) attention, we introduce a binary mask $\{C\} \in \{0, 1\}^{L\times L}$ defined by

$$C_{i,j} = \begin{cases} 1, & j \leq i, \\ 0, & j > i, \end{cases} \quad (8)$$

where: $C_{i,j} = 1$ allows position i to attend to

Table 1: **Compare results of our method with other SOTA methods on POPE, CHAIR, and MME datasets**. The best performances within each setting are **bolded**, baseline: LLaVA-1.5-7B.

Method	Venue	POPE		CHAIR				MME				
		F1↑	Acc↑	C_S↓	C_I↓	Recall	length	Exist.↑	Count↑	Pos.↑	Color↑	Total↑
Baseline	-	79.3	76.6	51.0	15.2	75.2	102.2	175.67	124.67	114.00	151.00	565.34
Dola (Chuang et al., 2023)	ICLR 2024	80.2	83.1	57.0	15.2	78.2	97.5	180.10	127.40	119.30	154.60	594.10
VCD (Leng et al., 2024)	CVPR 2024	85.3	85.0	51.0	14.9	77.2	101.9	184.66	137.33	128.67	153.00	603.66
OPERA (Huang et al., 2024)	CVPR 2024	84.2	85.2	47.0	14.6	78.5	95.3	180.67	133.33	111.67	123.33	549.00
DOPRA (Wei and Zhang, 2024)	MM 2024	84.6	84.3	46.3	13.8	78.2	96.1	185.67	138.33	120.67	133.00	577.67
HALC (Chen et al., 2024c)	ICML 2024	83.9	84.0	50.2	12.4	78.4	97.2	190.00	143.30	128.30	160.00	621.60
CCA-LLaVA (Xing et al., 2024a)	NeurIPS 2024	86.4	86.5	43.0	11.5	80.4	96.6	190.00	148.33	128.33	153.00	641.66
RITUAL (Woo et al., 2024)	Arxiv 2024	85.2	84.3	45.2	13.2	78.3	99.2	187.50	139.58	125.00	164.17	616.25
EAH (Zhang et al., 2024a)	EMNLP 2025	85.7	86.0	**36.4**	**9.9**	74.9	97.7	190.00	108.33	**145.00**	160.66	603.99
SID (Huo et al., 2025)	ICLR 2025	85.6	85.8	44.2	12.2	73.0	99.4	183.90	132.20	127.80	155.90	599.80
TAME (Tang et al., 2025a)	ICLR 2025	85.4	85.7	41.3	12.2	74.4	98.8	193.00	137.33	139.00	164.67	634.00
Vissink (Kang et al., 2025)	ICLR 2025	86.0	86.5	52.4	14.5	79.1	103.0	190.00	148.33	138.33	155.00	631.33
AGLA (An et al., 2024)	CVPR 2025	84.6	85.5	43.0	14.1	78.9	98.8	**195.00**	153.89	129.44	161.67	640.00
Farsight(Tang et al., 2025b)	CVPR 2025	-	-	41.6	13.2	75.5	100.6	-	-	-	-	-
ONLY (Wan et al., 2025)	ICCV 2025	85.5	85.1	49.8	14.3	75.9	99.7	191.67	145.55	136.66	161.66	635.55
MCA-LLaVA (Zhao et al., 2025)	MM 2025	86.0	86.5	38.0	10.9	76.6	92.5	190.00	**163.33**	126.67	170.00	650.00
Reverse-VLM (Wu et al., 2025b)	NeurIPS 2025	-	-	35.3	9.3	75.2	70.4	-	-	-	-	-
TokenTruth(ours)	-	**86.6**	**86.3**	38.0	11.7	75.2	98.2	**195.00**	156.66	126.66	**175.00**	**653.33**

position j (including itself and any past tokens), $C_{i,j} = 0$ blocks attention to future tokens ($j > i$), ensuring no peeking ahead. This mask is implemented as a lower triangular matrix (via torch.tril) and applied to the attention scores to enforce the autoregressive property.

We construct a position-aware penalty matrix $B^h \in \mathbb{R}^{n \times n}$ that implements the distance decay model from Equation 3:

$$B^h_{i,j} = -m_h \cdot |j - i|, \tag{9}$$

where m_h is the head-specific slope parameter. The slope parameters follow a geometric progression to ensure balanced attention redistribution across heads: for H heads, we compute the base slope $m = 2^{-8/H}$, then set $m_h = m^h$ for $h = 1, 2, ..., H$. When the number of heads is not a power of 2, additional slopes are computed through interpolation to maintain optimal coverage.

The modified attention score matrix becomes:

$$\mathbf{W} = \boldsymbol{A}^{\ell,h}_i + B^h_{i,j}, \tag{10}$$

inally, we obtain the optimized attention distribution that approximately solves Equation 6:

$$\tilde{\mathbf{W}} = \text{Softmax}(\mathbf{W}) * \mathbf{C}_{i,j}. \tag{11}$$

This design increases attention entropy $\Phi(\tilde{\boldsymbol{\alpha}}^{\ell,h}_i)$ while maintaining the constraints from our optimization framework, thereby reducing hallucination probability according to Equation 2.

5 Experiments

5.1 Experimental Setups

Baselines. To demonstrate the broad applicability of our method in LVLM architecture, we applied and evaluated the latest models, including LLaVA-1.5-7B (Liu et al., 2024a), Qwen2-VL-7B (Wang et al., 2024b). This study used the following data sets as evaluation sets, representing the expertise in reducing hallucination and general fields.

Evaluation Benchmarks. We conduct evaluations on image benchmarks. For image benchmarks, we assess three categories: (1) Comprehensive benchmarks (MMBench (Yuan et al., 2023), LLaVA^W (Liu et al., 2024a), MM-Vet (Yu et al., 2023), MME (Yin et al., 2023); (2) General VQA benchmarks (VizWiz (Gurari et al., 2018), ScienceQA (Lu et al., 2022); (3) Hallucination benchmarks (POPE(Li et al., 2023), CHAIR (Rohrbach et al., 2018)).

5.2 Evaluation Results on Hallucination Benchmarks

As shown in Table 1, the methods to mitigate hallucinations can be broadly classified into four groups. The first group, including OPERA (Huang et al., 2024), DOPRA (Wei and Zhang, 2024), DOLA(Chuang et al., 2023), VCD (Leng et al., 2024), HALC (Chen et al., 2024c), (An et al., 2024), RITUAL (Woo et al., 2024), AGLA (An et al., 2024), SID (Huo et al., 2025), Only (Wan et al., 2025), focuses on modifying the decoding process to address hallucinations. The second group, represented by LESS is more (Yue et al., 2024), adjusts the logits of the end-of-sequence (EOS) symbol to control its positioning, allowing the model to terminate earlier, thus reducing hallucinations. The third group, exemplified by CCA-LLaVA (Xing et al., 2024a), MCA-LLaVA (Zhao et al., 2025) and Reverse-VLM (Wu et al.,

Table 2: **Comparison of different LVLMs and TokenTruth across all image benchmarks**. Notably, in the Hallucination Benchmark, lower scores on $CHAIR_I$ and $CHAIR_S$ indicate better performance, while higher scores are preferable for other metrics.

Method	Comprehensive Benchmark			General VQA		Hallucination Benchmark				
	MMBench ↑	$LLaVA^W$	MM-Vet↑	VizWiz↑	SQA↑	$CHAIR_S$ ↓	$CHAIR_I$ ↓	POPE-R↑	POPE-F1↑	POPE-A↑
LLaVA-1.5-7B	64.3	72.5	30.5	48.5	65.5	48.0	13.9	87.0	85.4	84.0
+ICD	63.1	69.7	30.4	46.9	62.8	47.7	13.6	87.9	84.9	84.0
+VCD	63.9	70.9	29.5	43.4	63.3	46.8	13.2	87.0	85.3	85.0
+OPERA	64.4	72.0	31.4	50.0	64.9	45.2	12.7	88.8	84.2	85.2
+SID	65.0	73.4	31.2	50.9	67.8	44.2	14.0	89.4	85.6	85.8
+TAME	65.3	73.9	30.5	51.6	66.0	41.3	12.2	88.9	85.4	85.7
+Vissink	64.8	74.1	33.5	53.8	67.0	52.4	14.5	87.7	84.9	85.8
+TokenTruth	**65.3** (+1.0)	**74.8** (+2.3)	**33.8** (+3.3)	**54.3** (+5.8)	**66.5** (+1.0)	**39.4** (+8.6)	**11.2** (+2.7)	**90.6** (+3.6)	**86.6** (+1.2)	**86.3** (+2.3)
Qwen2-VL-7B	83.0	75.6	60.6	57.3	74.1	25.4	8.0	79.9	87.0	88.0
+TokenTruth	**83.3** (+0.3)	**76.8** (+1.2)	**62.2** (+1.6)	**58.8** (+1.5)	**74.7** (+0.6)	**23.6** (+1.8)	**7.8** (+0.2)	**80.3** (+0.4)	**87.3** (+0.3)	**88.2** (+0.2)
Qwen2.5-VL-7B	83.5	76.8	62.2	70.6	79.0	27.2	9.0	80.4	87.4	88.4
+TokenTruth	**84.3** (+0.8)	**77.9** (+1.1)	**64.8** (+2.6)	**71.3** (+0.7)	**80.8** (+1.8)	**23.0** (+4.2)	**8.5 (+0.5)**	**80.9 (+0.5)**	**87.8** (+0.4)	**88.7** (+0.3)

2025b), investigates the weakening of the information flow between the visual and the instruction tokens caused by the long-term decay in rotational position encoding (RoPE). To address this, they proposed Concentric Causal Attention (CCA), which reorganizes the positions of visual tokens and corrects the causal mask to alleviate object hallucinations. The fourth group includes EAH (Zhang et al., 2024a), Vissink (Kang et al., 2025), Farsight (Tang et al., 2025b)and ONLY (Wan et al., 2025), which aim to enhance the truthfulness of the model's output during inference by adjusting attention.

Among these methods, TokenTruth shows competitive performance and achieves notable results. TokenTruth achieves state-of-the-art (SOTA) performance in POPE and certain VQA tasks that test fine-grainedness, as evenly distributing visual attention to other tokens improves the model's global attention to the image. TokenTruth is second only to EAH on descriptive datasets such as CHAIR. Compared with EAH's approach of directly replacing the attention head, TokenTruth is more reasonable because it does not change the internal representation of the model. Therefore, in terms of recall, TokenTruth leads other methods and reaches 82.1. These results show that TokenTruth effectively targets fine-grained object-related issues in the hallucination dataset and achieves satisfactory results.

5.3 Evaluation Results of TokenTruth on General Vision-Language Tasks and Benchmarks

As demonstrated in Table 1, our proposed TokenTruth method achieves significant and consistent

Table 3: Performance comparison of different models with sigma augmentation on various benchmarks.

Model	MMBench ↑	$LLaVA^W$↑	MM-Vet ↑	VizWiz ↑	SQA ↑
LlaVA-1.5-7B	64.3	72.5	30.5	48.5	65.5
+σ-0.9	65.3	74.8	33.8	54.3	66.5
+σ-0.8	65.2	74.3	33.4	53.7	66.1
+σ-0.7	64.7	73.7	32.8	52.9	65.8
Qwen2-VL-7B	83.0	75.6	60.6	57.3	74.1
+σ-0.9	83.1	76.4	61.9	58.2	74.4
+σ-0.8	83.3	76.8	62.2	58.8	74.7
+σ-0.7	83.0	76.1	61.3	57.9	74.2
Qwen2.5-VL-7B	83.5	76.8	62.2	70.6	79.0
+σ-0.9	84.1	77.4	64.1	70.9	80.2
+σ-0.8	84.3	77.9	64.8	71.3	80.8
+σ-0.7	83.7	77	63.5	70.7	79.3

performance gains in all benchmark datasets compared to the baseline model, LLaVA-1.5. In particular, MME evaluates fundamental perception capabilities, such as object existence, counting, and attribute recognition (e.g., color and position).

The observed enhancements underscore TokenTruth's ability to strengthen general visual perception in large vision language models (LVLM) by dynamically reallocating attention to critical visual tokens. This not only mitigates visual attention misalignment, but also fosters a more robust and accurate multimodal understanding. Consistent gains in diverse datasets further validate the versatility and generalizability of TokenTruth.

Generalization Across VQA Benchmarks and Models

In addition, TokenTruth delivers consistent performance improvements across multiple visual question-answering (VQA) benchmarks, demonstrating its robustness in tasks requiring precise visual feature localization and interpretation. This aligns with the core mechanism of TokenTruth,

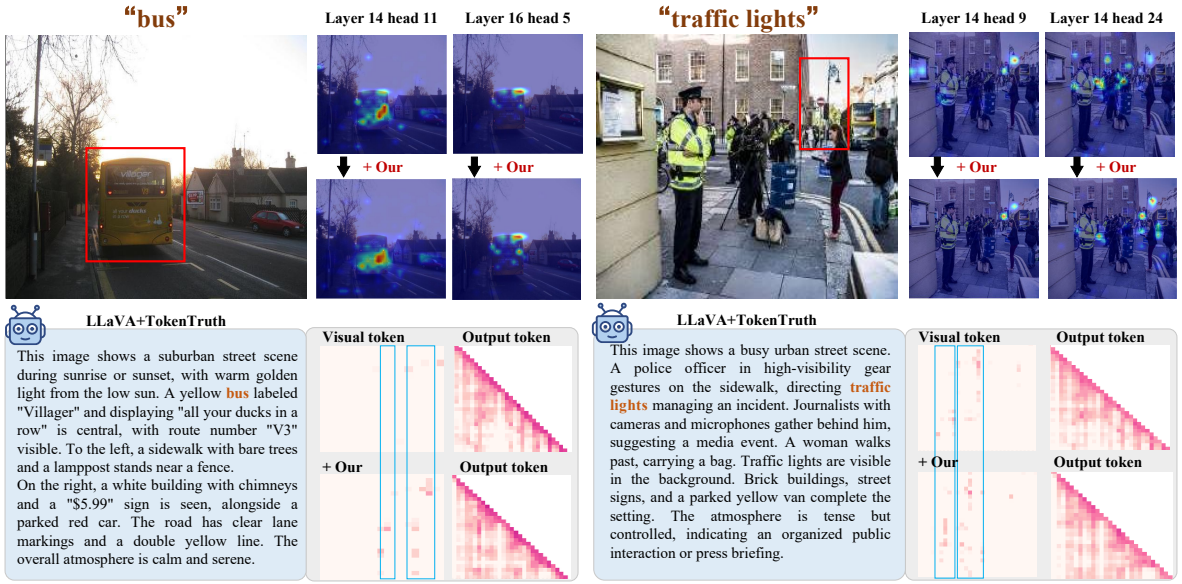

Figure 6: **The case study of TokenTruth**. It can seen that some of the original cases of object hallucination have been mitigated by the TokenTruth, baseline: LLaVA-1.5-7B.

redirecting attention to semantically relevant visual tokens, which are crucial to VQA accuracy.

To further verify the generalizability of TokenTruth, we applied it to Qwen2-VL, another prominent LVLM. As evidenced in Table 2, we observe that Qwen2-VL also exhibits the phenomenon of the attentional sink, which reinforces the broader relevance of this issue. After integrating TokenTruth, Qwen2-VL achieves measurable performance gains, particularly in VQA tasks, confirming that TokenTruth benefits extend beyond a single model architecture.

5.4 Ablation Study of TokenTruth

As shown in Figure 6, token-level penalty significantly improves the model's ability to recognize visual content and generate coherent descriptions. For example, while LLaVA-1.5 may hallucinate "*a man is standing on a clothesline*," our TokenTruth method correctly identifies the scene as "*a man is standing on a folding table*," effectively reducing initial errors.

This token-level correction supports more accurate and semantically consistent follow-up descriptions—like "he is ironing clothes" minimizing the risk of compounding hallucinations. The refined attention mechanism enhances both object recognition and the richness of activity descriptions through better modeling of object interactions.

Additionally, head-level enhancement improves contextual integration. For instance, generating "*which is attached to a yellow car, and he is ironing clothes*" shows stronger environmental grounding than baseline models, indicating improved semantic precision and contextual awareness in vision-language generation.

5.5 Ablation Study on Global Decady σ

we have conducted a comprehensive ablation study on the hyperparameter σ across multiple models (LlaVA-1.5-7B, Qwen2-VL-7B, and Qwen2.5-VL-7B) and key benchmarks. As shown in Table 3, we evaluate $\sigma \in \{0.7, 0.8, 0.9\}$ and observe consistent trends:

These results confirm that $\sigma = 0.9$ or $\sigma = 0.8$ that LLaVA1.5/Qwen2-VL offer the optimal trade-off between preserving content relevance and suppressing hallucinatory tokens. We have added this analysis to Section 4.3 to strengthen the empirical justification of our design choice.

6 Conclusion

This work introduces TokenTruth, a novel strategy that effectively reduces hallucinations in LVLMs by dynamically adjusting attention to prioritize meaningful visual content. By addressing key issues like excessive activation tokens and positional information decay, the method enhances contextual coherence and maintains fluent, relevant generation. Our method effectively reduces hallucinations while maintaining fluency and relevance in vision-language generation tasks.

7 Limitation

While TokenTruth effectively mitigates hallucinations caused by attention sinks and positional decay in LVLMs, several limitations remain. Our token-level penalty mechanism relies on identifying "critical" tokens via activation magnitudes in query/key projections, which may not fully capture semantic importance in all modalities or tasks—especially when visual inputs are ambiguous or highly abstract. Finally, while we evaluate on standard image understanding benchmarks, real-world web-scale applications involve more complex multimodal inputs (e.g., multi-image, video, or user-contextualized prompts), where dynamic token suppression may require adaptive thresholds or task-aware scheduling—directions we leave for future work.

References

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

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. 2024. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*.

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. 2025. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. *arXiv preprint arXiv:2503.15871*.

Jiaqi Bai, Hongcheng Guo, Zhongyuan Peng, Jian Yang, Zhoujun Li, Mohan Li, and Zhihong Tian. 2025. Mitigating hallucinations in large vision-language models by adaptively constraining information flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23442–23450.

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

Jinhe Bi, Yifan Wang, Danqi Yan, Aniri, Wenke Huang, Zengjie Jin, Xiaowen Ma, Artur Hecker, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. 2025a. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *Preprint*, arXiv:2502.12119.

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025b. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15230–15250, Vienna, Austria. Association for Computational Linguistics.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. 2025. Eazy: Eliminating hallucinations in lvlms by zeroing out hallucinatory image tokens. *arXiv preprint arXiv:2503.07772*.

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

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2025a. Attention hijackers: Detect and disentangle attention hijacking in lvlms for hallucination mitigation. *arXiv preprint arXiv:2503.08216*.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025b. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024b. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *18th European Conference on Computer Vision ECCV 2024*.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*.

Peng Gao, Yujian Lee, Xiaofeng Zhang, Zailong Chen, and Hui Zhang. 2025. Remember me: Bridging the

long-range gap in lvlms with three-step inference-only decay resilience strategies. *arXiv preprint arXiv:2511.09868*.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024a. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024b. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*.

Jihao Gu, Yingyao Wang, Meng Cao, Pi Bu, Jun Song, Yancheng He, Shilong Li, and Bo Zheng. 2024. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation. *arXiv preprint arXiv:2412.14487*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2024. Cracking the code of hallucination in lvlms with vision-aware head divergence. *arXiv preprint arXiv:2412.13949*.

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

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

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024a. 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.

Kailin Jiang, Hongbo Jiang, Ning Jiang, Zhi Gao, Jinhe Bi, Yuchen Ren, Bin Li, Yuntao Du, Lei Liu, and Qing Li. 2025a. Kore: Enhancing knowledge injection for large multimodal models via knowledge-oriented augmentations and constraints. *Preprint*, arXiv:2510.19316.

Kailin Jiang, Ning Jiang, Yuntao Du, Yuchen Ren, Yuchen Li, Yifan Gao, Jinhe Bi, Yunpu Ma, Qingqing Liu, Xianhao Wang, Yifan Jia, Hongbo Jiang, Yaocong Hu, Bin Li, and Lei Liu. 2025b. Mined: Probing and updating with multimodal time-sensitive knowledge for large multimodal models. *Preprint*, arXiv:2510.19457.

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

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. 2025. Massive values in self-attention modules are the key to contextual knowledge understanding. *arXiv preprint arXiv:2502.01563*.

Zhehan Kan, Ce Zhang, Zihan Liao, Yapeng Tian, Wenming Yang, Junyuan Xiao, Xu Li, Dongmei Jiang, Yaowei Wang, and Qingmin Liao. 2024. Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms. *arXiv preprint arXiv:2411.12713*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.

Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *arXiv preprint arXiv:2406.01920*.

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

Bin Li, Dehong Gao, Yeyuan Wang, Linbo Jin, Shanqing Yu, Xiaoyan Cai, and Libin Yang. 2025a. Instruction-aligned visual attention for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2503.18556*.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. 2025b. Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. *arXiv preprint arXiv:2501.01926*.

Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. 2025c. Treble counterfactual vlms: A causal approach to hallucination. *arXiv preprint arXiv:2503.06169*.

Shuo Li, Jiajun Sun, Guodong Zheng, Xiaoran Fan, Yujiong Shen, Yi Lu, Zhiheng Xi, Yuming Yang, Wenming Tan, Tao Ji, et al. 2025d. Mitigating object hallucinations in mllms via multi-frequency perturbations. *arXiv preprint arXiv:2503.14895*.

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

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. 2025e. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*.

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

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Maria Lymperaiou, Giorgos FIlandrianos, Angeliki Dimitriou, Athanasios Voulodimos, and Giorgos Stamou. 2025. Halcece: A framework for explainable hallucination detection through conceptual counterfactuals in image captioning. *arXiv preprint arXiv:2503.00436*.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13151–13160.

Shunqi Mao, Chaoyi Zhang, and Weidong Cai. 2025. Through the magnifying glass: Adaptive perception magnification for hallucination-free vlm decoding. *arXiv preprint arXiv:2503.10183*.

Dexter Neo and Tsuhan Chen. 2024. Vord: Visual ordinal calibration for mitigating object hallucinations in large vision-language models. *arXiv preprint arXiv:2412.15739*.

Tianfan Peng, Yuntao Du, Pengzhou Ji, Shijie Dong, Kailin Jiang, Mingchuan Ma, Yijun Tian, Jinhe Bi, Qian Li, Wei Du, Feng Xiao, and Lizhen Cui. 2025. Can visual input be compressed? a visual token compression benchmark for large multimodal models. *Preprint*, arXiv:2511.02650.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Xuankun Rong, Wenke Huang, Jian Liang, Jinhe Bi, Xun Xiao, Yiming Li, Bo Du, and Mang Ye. 2025. Backdoor cleaning without external guidance in mllm fine-tuning. *arXiv preprint arXiv:2505.16916*.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*.

Yuying Shang, Xinyi Zeng, Yutao Zhu, Xiao Yang, Zhengwei Fang, Jingyuan Zhang, Jiawei Chen, Zinan Liu, and Yu Tian. 2024. From pixels to tokens: Revisiting object hallucinations in large vision-language models. *arXiv preprint arXiv:2410.06795*.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. *arXiv preprint arXiv:2503.00361*.

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

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. 2025b. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. 2025. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. *arXiv preprint arXiv:2507.00898*.

Chao Wang, Weiwei Fu, and Yang Zhou. 2025a. Tpc: Cross-temporal prediction connection for vision-language model hallucination reduction. *arXiv preprint arXiv:2503.04457*.

Jiaqi Wang, Yifei Gao, and Jitao Sang. 2024a. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023a. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yujun Wang, Jinhe Bi, Yunpu Ma, and Soeren Pirk. 2025b. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. *arXiv preprint arXiv:2506.14766*.

Zehao Wang, Xinpeng Liu, Xiaoqian Wu, Yudonglin Zhang, Zhou Fang, Yifan Fang, Junfu Pu, Cewu Lu, and Yong-Lu Li. 2024c. Verb mirage: Unveiling and assessing verb concept hallucinations in multimodal large language models. *arXiv preprint arXiv:2412.04939*.

Jinfeng Wei and Xiaofeng Zhang. 2024. Dopra: Decoding over-accumulation penalization and re-allocation in specific weighting layer. *Proceedings of the 32nd ACM International Conference on Multimedia*.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2024. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*.

Jiarui Wu, Zhuo Liu, and Hangfeng He. 2025a. Mitigating hallucinations in multimodal spatial relations through constraint-aware prompting. *arXiv preprint arXiv:2502.08317*.

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E Gonzalez, Trevor Darrell, and David M Chan. 2025b. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. *arXiv preprint arXiv:2504.13169*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.

Chunzhao Xie, Tongxuan Liu, Lei Jiang, Yuting Zeng, Yunheng Shen, Weizhe Huang, Jing Li, Xiaohua Xu, et al. 2025. Tarac: Mitigating hallucination in lvlms via temporal attention real-time accumulative connection. *arXiv preprint arXiv:2504.04099*.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024a. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024b. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2024. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. *arXiv preprint arXiv:2412.13817*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2503.13107*.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. 2024. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*.

Liu Yuan, Duan Haodong, Zhang Yuanhan, Li Bo, Zhang Songyang, and Zhao Wangbo. 2023. Mmbench:is your multi-modal model an all-around player. *arXiv:2307.06281*.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *The 62nd Annual Meeting of the Association for Computational Linguistics*.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. 2025a. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2502.06130*.

Gengyuan Zhang, Jinhe Bi, Jindong Gu, Yanyu Chen, and Volker Tresp. 2023. Spot! revisiting video-language models for event understanding. *arXiv preprint arXiv:2311.12919*.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024a. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*.

Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2024b. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv preprint arXiv:2406.06579*.

Xiaofeng Zhang, Fanshuo Zeng, and Chaochen Gu. 2024c. Simignore: Exploring and enhancing multimodal large model complex reasoning via similarity computation. *Neural Networks*, page 107059.

Xiaofeng Zhang, Fanshuo Zeng, and Chaochen Gu. 2025b. Simignore: Exploring and enhancing multimodal large model complex reasoning via similarity computation. *Neural Networks*, 184:107059.

Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. 2024d. Enhancing multimodal large language models complex reason via similarity computation. *arXiv preprint arXiv:2412.09817*.

Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. 2025c. Enhancing multimodal large language models complex reason via similarity computation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Jiawei Cao, Hao Cheng, and Kaijie Wu. 2026a. What drives attention sinks? a study of massive activations and rotational positional encoding in large vision–language models. *Information Processing & Management*, 63(2):104431.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, et al. 2026b. Hallucination begins where saliency drops. In *The Fourteenth International Conference on Learning Representations*.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Mingjia Zhang, and Baobao Chang. 2024. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. *arXiv preprint arXiv:2411.14279*.

Qiyan Zhao, Xiaofeng Zhang, Shuochen Chang, Qianyu Chen, Xiaosong Yuan, Xuhang Chen, Luoqi Liu, Jiajun Zhang, Xu-Yao Zhang, and Da-Han Wang. 2026. Context tokens are anchors: Understanding the repetition curse in dmllms from an information flow perspective. In *The Fourteenth International Conference on Learning Representations*.

Qiyan Zhao, Xiaofeng Zhang, yiheng Li, Yun Xing, Yuan Xiaosong, Feilong Tang, Sinan Fan, Xuhang Chen, Xuyao Zhang, and Dahan Wang. 2025. Mcallava: Manhattan causal attention for reducing hallucination in large vision-language models. *The 33rd ACM International Conference on Multimedia*.

Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. 2024. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *arXiv preprint arXiv:2411.12591*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2025. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations*.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*.

Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.