# ATTEND TO VISION WHEN IT MATTERS: VISUAL DEPENDENCY-AWARE ATTENTION REBALANCING FOR LARGE VISION-LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Vision-Language Models (LVLMs) commonly suffer from visual hallucinations, generating content that deviates from visual evidence. This problem stems from the dominance of linguistic priors in autoregressive generation and the insufficient visual grounding. Existing inference-time attention interventions mitigate hallucinations by strengthening visual attention, but typically rely on native attention patterns and adopt static or uniform enhancement strategies, failing to distinguish informative visual tokens, identify layers that benefit most from intervention, or adapt to the dynamic demand for visual evidence during generation. Such indiscriminate intervention may alleviate hallucinations at the cost of language generation quality. To address this, we propose ViDRA, a fine-grained attention rebalancing framework. During prefill, ViDRA estimates visual token importance by jointly considering representation stability and instruction relevance, while deriving layer-wise intervention strength from visual activity across Transformer layers. During decoding, a lightweight Visual Dependency Gate predicts the visual dependency of the upcoming token, enabling visual attention to be enhanced only when needed. Extensive experiments on visual hallucination benchmarks and general vision-language tasks show that ViDRA effectively reduces visual hallucinations while preserving language generation quality and inference efficiency, achieving state-of-the-art performance across multiple LVLM backbones.

## 1 INTRODUCTION

Large Vision-Language Models (LVLMs), which integrate visual encoders with large language models, demonstrate strong cross-modal understanding across diverse vision-language tasks (Li et al., 2024; Lai et al., 2024; Li et al., 2026). Despite these advances, LVLMs often underutilize visual evidence and over-rely on linguistic priors during autoregressive generation (Liu et al., 2024d; Leng et al., 2024), resulting in visual hallucinations such as entity and attribute errors. Therefore, effectively leveraging visual information while preserving the model's intrinsic language generation capability is crucial for improving the reliability of LVLMs.

To mitigate visual hallucinations, recent studies have explored inference-time attention interventions to strengthen visual grounding. PAI (Liu et al., 2024d) alleviates language-dominant bias by amplifying attention to image tokens. VAR (Kang et al., 2025) identifies the visual attention sink and redistributes excessive attention from anchor tokens to other visual tokens. VisFlow (Tang et al., 2026) further models the unequal contributions of visual tokens and attention heads, selectively enhancing informative visual tokens while suppressing abnormal textual attention. Collectively, these methods demonstrate that attention intervention is an effective and deployment-efficient strategy for mitigating visual hallucinations without retraining LVLMs.

Despite these advances, existing attention intervention methods still struggle to balance visual attention enhancement with the preservation of original language generation capability, as shown in Fig. 1. First, visual tokens differ inherently in enhancement value. Uniformly amplifying all visual tokens (Liu et al., 2024d) exacerbates the attention sink by widening the attention gap between anchor and ordinary visual tokens. Although subsequent methods (Tang et al., 2026; Che et al., 2025) adopt token-wise selective enhancement, their intervention criteria still rely on native attention



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

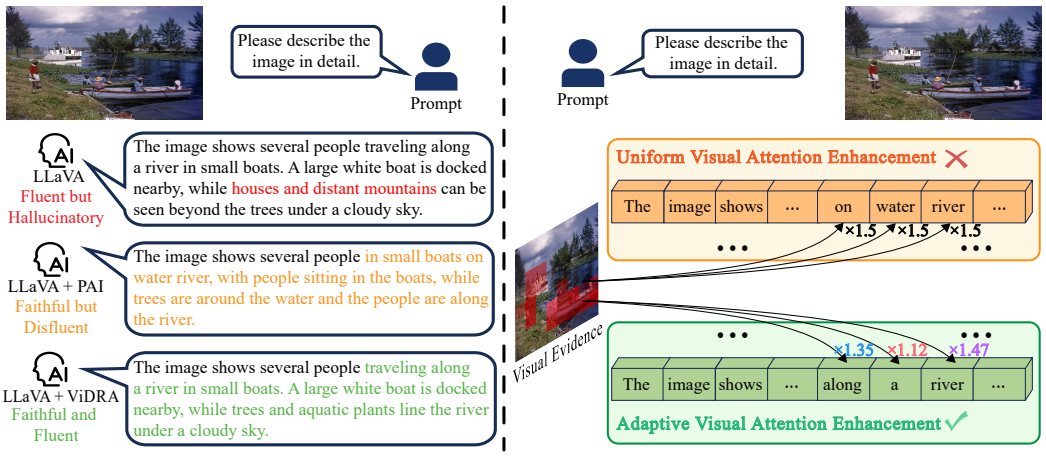



Figure 1: Motivation of our ViDRA. Left: Uniform visual attention enhancement (PAI) mitigates hallucinations but degrades fluency, while ViDRA generates faithful and fluent responses. Right: Unlike uniform enhancement, ViDRA adaptively modulates visual attention based on the visual dependency of each generation step.

patterns and thus remain susceptible to inherent attention biases. Second, visual dependency varies dynamically across generated tokens: entity, attribute, and quantity tokens often rely heavily on visual evidence, whereas connectives and other syntactic tokens rely primarily on linguistic context. Existing methods (Kang et al., 2025; Tang et al., 2026) overlook this variation and apply static interventions throughout generation, introducing irrelevant visual noise and disrupting linguistic reasoning. Therefore, mitigating visual hallucinations requires not merely increasing visual attention but selectively enhancing critical visual evidence at appropriate generation steps.

Motivated by these, we propose ViDRA (Visual Dependency-Aware Rebalancing of Attention) that adaptively enhances visual attention within the standard prefill–decoding pipeline of LVLMs along three complementary dimensions: visual tokens, Transformer layers, and generation steps. During prefill, ViDRA estimates visual token importance via foreground confidence and instruction relevance, while measuring layer-wise visual activity to identify where intervention is most beneficial. During decoding, a lightweight Visual Dependency Gate predicts the visual dependency of the upcoming token, dynamically determining when and how strongly visual attention should be enhanced. The gate is supervised by prediction discrepancies between full-vision and vision-drop conditions and is trained independently of the LVLM backbone. Without modifying the backbone or pretrained weights, ViDRA selectively reinforces visual evidence while preserving intrinsic language generation capability, thereby effectively mitigating visual hallucinations.

Our contributions are summarized as follows: (i) We reveal the dynamic visual dependency during autoregressive generation, showing that visual demand varies substantially across generation steps while native visual attention fails to reflect this dependency. (ii) We propose ViDRA, a fine-grained attention intervention framework that jointly models visual token importance, layer-wise visual activity, and step-wise visual dependency. (iii) Extensive experiments on hallucination benchmarks and general vision-language tasks demonstrate that ViDRA reduces visual hallucinations while preserving generation quality and inference efficiency, outperforming state-of-the-art methods.

## 2 Related Work

**Large Vision-Language Models.** Existing LVLMs typically follow a three-stage architecture: visual encoder, projector, and language model. They differ mainly in their cross-modal alignment designs. The LLaVA family (Liu et al., 2023; 2024b) uses linear projections or lightweight MLPs, while BLIP-2 (Li et al., 2023a), MiniGPT-4 (Zhu et al., 2024), and InstructBLIP (Dai et al., 2023) adopt query-based modules such as Q-Former to extract text-relevant visual information. Subsequent works including Qwen-VL (Bai et al., 2023; Wang et al., 2024a), mPLUG-Owl2 (Ye et al., 2024), InternVL (Chen et al., 2024b;a; Zhu et al., 2025; Wang et al., 2025), and LLaVA-NeXT (Liu et al., 2024c)

further improve multimodal understanding via stronger visual encoders, high-resolution inputs, large-scale pretraining, and instruction tuning. Despite these advances, LVLMs remain susceptible to visual hallucinations (Rohrbach et al., 2018; Li et al., 2023b; Sun et al., 2024).

**Visual Hallucination Mitigation.** Existing methods for hallucination mitigation can be broadly divided into training-based optimization and inference-time intervention. Training-based methods improve cross-modal alignment via hallucination-aware supervision (Liu et al., 2024a; Jiang et al., 2024) or preference optimization, such as RLHF (Yu et al., 2024) and DPO (Yang et al., 2025), but typically incur high training costs. Inference-time methods avoid backbone retraining and mainly include contrastive decoding, post-processing correction, and attention intervention. VCD (Leng et al., 2024) and ICD (Wang et al., 2024b) suppress linguistic priors by constructing contrastive branches via input perturbations, at the cost of additional forward passes and potential noise. Post-processing methods such as Woodpecker (Yin et al., 2024) use external modules to detect and correct hallucinated content, resulting in longer inference pipelines. The most relevant to our work are the attention intervention methods. PAI (Liu et al., 2024d), VAR (Kang et al., 2025), and VisFlow (Tang et al., 2026) enhance visual grounding through global visual enhancement, attention sink redistribution, and selective token/head intervention, respectively. EAZY (Che et al., 2025) suppresses hallucinatory image tokens, while ASCD (Wang et al., 2026a) identifies text-centric attention heads and critical visual tokens for attention steering and contrastive decoding. Despite increasingly fine-grained designs, existing methods still rely heavily on native attention patterns and apply static strategies across generation steps, overlooking dynamic visual dependency during generation.

## 3 Empirical Analysis and Motivation

Existing studies show that visual attention enhancement mitigates visual hallucination, but indiscriminate amplification impairs language generation. This raises a core question: which visual tokens should be prioritized, and when should visual attention be enhanced during generation? To answer this, we analyze the visual token representation and the dynamic visual dependency during generation. We obtain two key observations: (1) foreground visual tokens exhibit greater representation stability under low-pass perturbations along the feature dimension, providing an intrinsic cue to identify informative visual regions; and (2) visual dependency varies substantially across generated tokens, but the model's native visual attention fails to capture this dependency.

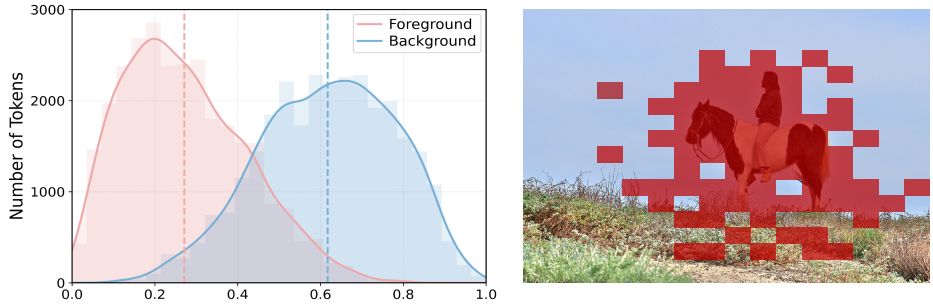
### 3.1 Channel-Wise Low-Frequency Stability of Foreground Tokens

Existing attention interventions (Liu et al., 2024d; Tang et al., 2026) typically estimate visual token importance directly from native attention weights, which can be affected by positional bias (Wang et al., 2026c), attention sink (Kang et al., 2025), and other factors. Thus, high attention on a visual token does not necessarily correspond to high-value visual information. Instead of relying solely on attention magnitude, we investigate whether informative foreground tokens can be identified from the intrinsic properties of their visual representations.

LaSt-ViT (Shi et al., 2026) shows that the foreground regions in natural images tend to exhibit more homogeneous semantics and smoother channel-wise feature variations, whereas the background regions are more semantically diverse and exhibit stronger fluctuations across channels. Inspired by this, we exploit the representation stability under channel-wise low-pass perturbation as an intrinsic cue for identifying informative foreground tokens. Given the $i$-th visual token $x_i \in \mathbb{R}^D$, we apply a 1D discrete Fourier transform (DFT) along the feature dimension: $X_i = \mathcal{F}(x_i)$. We then retain the low-frequency components using a low-pass filter $H_{\mathrm{LP}}$ and reconstruct the token representation via the inverse transform: $\hat{x}_i = \mathcal{F}^{-1}(H_{\mathrm{LP}} \odot X_i)$. The low-pass reconstruction error is defined as:

$$d_i = \frac{1}{D} \sum_{j=1}^{D} \frac{|\hat{x}_{i,j} - x_{i,j}|}{|x_{i,j}| + \epsilon}, \tag{1}$$

where a smaller $d_i$ indicates greater representation stability under channel-wise low-pass perturbation. Using ground-truth masks, we partition visual tokens into foreground and background sets, denoted $\mathcal{T}_{\mathrm{fg}}$ and $\mathcal{T}_{\mathrm{bg}}$, respectively. As shown in Fig. 2, we observe $\mathbb{E}_{i \in \mathcal{T}_{\mathrm{fg}}}[d_i] < \mathbb{E}_{i \in \mathcal{T}_{\mathrm{bg}}}[d_i]$, indicating that foreground tokens are more stable under low-pass perturbations, whereas background tokens exhibit stronger high-frequency variations. When mapped back to the image space, tokens with low reconstruction error predominantly lie within foreground regions.




(a) Distribution of $d_i$ for FG and BG Tokens (b) Spatial Distribution of Low-$d_i$ Tokens

Figure 2: Analysis of channel-wise low-pass reconstruction error. Evaluated on 100 randomly selected COCO (Lin et al., 2014) images with corresponding ground-truth masks.

These observations indicate that low-frequency stability provides an intrinsic cue for identifying informative foreground tokens. Unlike attention-based criteria, this signal is derived from visual representations and is less susceptible to inherent attention biases. However, foreground regions are not necessarily relevant to the instruction. When answering "What color is the car?", pedestrians and buildings also belong to the foreground but are irrelevant to the query. Therefore, in Sec. 4.2, we combine the foreground prior with instruction relevance to estimate visual token importance.

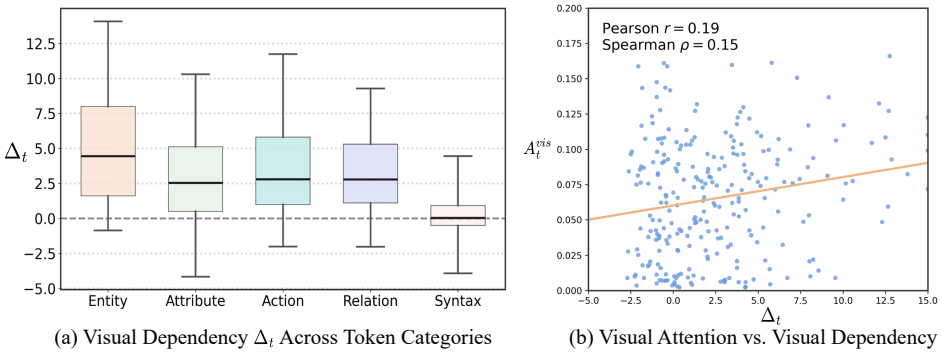
### 3.2 Dynamic Visual Dependency During Autoregressive Generation

LVLMs generate responses autoregressively and different generated tokens exhibit distinct visual dependencies. Object, attribute, and quantity tokens typically rely heavily on visual evidence, whereas articles and conjunctions depend primarily on linguistic context. To quantify the visual dependency of the target token $y_t$ at generation step $t$, we compare its prediction probability under full-vision and vision-drop conditions, and define the visual dependency score as the log-probability ratio:

$$p_{\text{full},t} = p(y_t \mid I, Q, y_{<t}), \quad p_{\text{drop},t} = p(y_t \mid \emptyset, Q, y_{<t}), \tag{2}$$

$$\Delta_t = \log p_{\text{full},t} - \log p_{\text{drop},t} = \log \frac{p_{\text{full},t}}{p_{\text{drop},t}}, \tag{3}$$

where $I$ and $Q$ denote the input image and the textual instruction, respectively. A larger $\Delta_t$ indicates a stronger reliance on visual evidence when predicting $y_t$.




(a) Visual Dependency $\Delta_t$ Across Token Categories (b) Visual Attention vs. Visual Dependency

Figure 3: Analysis of dynamic visual dependency. (a) Content-bearing tokens exhibit stronger visual dependency than syntactic tokens. (b) Visual attention is weakly correlated with visual dependency.

We categorize the generated tokens into Entity, Attribute, Action, Relation, and Syntax, and compare their $\Delta_t$ distributions. As shown in Fig. 3 (a), Entity, Attribute, and Action tokens exhibit higher visual dependency, while Syntax tokens rely less on visual evidence. However, the distributions overlap substantially and show considerable intra-category variation. This indicates that semantic categories capture only coarse trends and cannot reliably determine token-level visual dependency,

since visual dependency is inherently context-dependent. For example, generating "car" for "What color is the car?" requires little visual evidence ($\Delta_t = 0.12$), as the object is already specified in the query. In contrast, predicting "car" for "What is in front of the house?" requires visual object identification and exhibits much stronger visual dependency ($\Delta_t = 11.38$). Thus, even the same token can exhibit substantially different visual dependencies in different multimodal contexts.

We further compute the visual attention mass $A_t^{\text{vis}}$ allocated to visual tokens at each generation step $t$ and compare it with $\Delta_t$. As shown in Fig. 3 (b), native visual attention is only weakly correlated with visual dependency, with frequent attention–dependency mismatches: some tokens receive high visual attention despite low visual dependency, while others with strong visual dependency receive insufficient attention. This indicates that visual attention cannot reliably reflect the actual demand for visual evidence. The visual dependency of the upcoming token should be inferred from the current multimodal context. Consequently, we introduce the Visual Dependency Gate in Sec. 4.3.

# 4 METHOD

## 4.1 OVERVIEW

Given an input image $I$ and textual instruction $Q$, LVLMs encode the image into visual tokens $\mathcal{T}_{\text{vis}} = \{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^D$ and process them jointly with textual tokens through $L$ Transformer decoder layers to generate responses. LVLMs inference comprises two stages: prefill processes the complete multimodal input to establish context and initialize the KV cache, while decoding generates responses autoregressively by reusing the cached states. This pipeline naturally enables fine-grained visual attention intervention. Visual token importance and layer-wise visual activity can be estimated once during prefill and reused throughout generation, whereas visual dependency varies across generated tokens and must be modeled during decoding. We thus propose ViDRA (Visual Dependency-Aware Rebalancing of Attention), which adaptively rebalances visual attention along three complementary dimensions: visual tokens, Transformer layers, and generation steps.

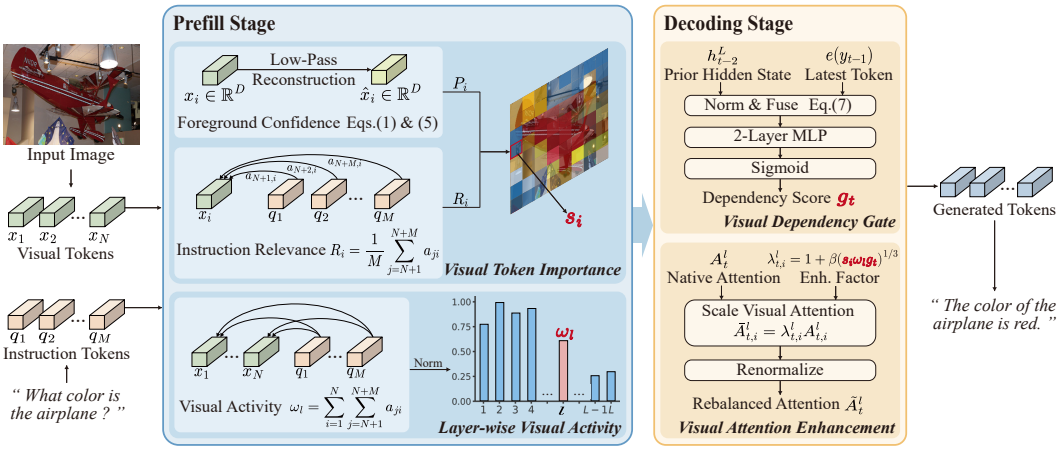



Figure 4: Overall architecture of the proposed ViDRA method.

As illustrated in Fig. 4, during the prefill stage, ViDRA computes visual token importance $s_i$ and layer-wise visual activity $\omega_l$, determining *"which visual tokens to enhance"* and *"which layers to intervene in"*, respectively. During decoding, Visual Dependency Gate predicts the visual dependency score $g_t \in [0, 1]$ of the upcoming token, determining *"when and how strongly to enhance visual attention"*. Specifically, for the $i$-th visual token at the $t$-th generation step and the $l$-th Transformer layer, we define the enhancement factor as $\lambda_{t,i}^l = 1 + \beta\,(s_i \omega_l g_t)^{1/3}$, where $\beta$ is a global enhancement coefficient. Given the original attention weight $A_{t,i}^l$, ViDRA enhances visual attention as:

$$\bar{A}_{t,i}^l = \lambda_{t,i}^l A_{t,i}^l, \qquad i \in \mathcal{T}_{\text{vis}}. \tag{4}$$

Non-visual attention is left unscaled, and the final rebalanced distribution $\tilde{A}_{t,i}^l$ is obtained by renormalization over all tokens.

### 4.2 Prefill Stage: Estimating Visual Intervention Priors

**Visual Token Importance.** As shown in Sec. 3.1, foreground visual tokens exhibit greater representation stability under low-pass perturbation. Based on the low-pass reconstruction error $d_i$ in Eq. 1, we apply min-max normalization across visual tokens to obtain foreground confidence:

$$P_i = 1 - \mathrm{MinMax}(d_i), \qquad P_i \in [0, 1]. \tag{5}$$

A larger $P_i$ indicates greater low-frequency stability and higher foreground confidence. However, foreground tokens are not necessarily relevant to the instruction. We therefore estimate instruction relevance $R_i \in [0, 1]$ from the attention received by each visual token from the instruction. The final visual token importance is defined as:

$$s_i = \alpha P_i + (1 - \alpha) R_i, \qquad s_i \in [0, 1], \tag{6}$$

where $\alpha$ is a weighting coefficient. A large $s_i$ indicates that the visual token is both informative and instruction-relevant, and thus receives stronger attention enhancement.

**Layer-wise Visual Activity.** Prior studies (Liu et al., 2024d; Zou et al., 2024; Che et al., 2025) have revealed that visual engagement varies substantially across Transformer layers. Shallow layers are dominated by cross-modal interaction and exhibit high visual activity, whereas deeper layers increasingly focus on high-level semantic and linguistic reasoning with decaying visual attention. This suggests that uniformly enhancing visual attention across all layers is suboptimal. We therefore prioritize visually active layers, where intervention is more likely to reinforce visual evidence while minimizing disruption to language generation. Specifically, we compute the total attention allocated to all visual tokens at the $l$-th layer and normalize it across layers to obtain the visual activity score $\omega_l \in [0, 1]$. A larger $\omega_l$ indicates that visual intervention yields greater benefits at that layer.

### 4.3 Decoding Stage: Visual Dependency Gating

Visual token importance $s_i$ and layer-wise visual activity $\omega_l$ determine where visual attention should be enhanced across tokens and layers, but neither captures whether the upcoming token actually requires visual evidence. As shown in Sec. 3.2, visual dependency varies dynamically with the multimodal context. We thus introduce the Visual Dependency Gate during decoding to predict the visual dependency of the upcoming token before each Transformer forward pass.

The gate takes two inputs: the hidden state $h^L_{t-2}$ from the prior decoding step, which summarizes the preceding multimodal context, and the embedding $e(y_{t-1})$ of the latest generated token, which provides the most recent semantic information. Note that $h^L_{t-1}$ is unavailable before the current Transformer forward pass, as it is produced only after processing $y_{t-1}$ and is subsequently used to predict $y_t$. The two inputs are normalized and fused into $z_t$, which is then fed into a lightweight bottleneck MLP with a sigmoid output:

$$z_t = \mathrm{Norm}(h^L_{t-2}) + \gamma \cdot \mathrm{Norm}(e(y_{t-1})), \tag{7}$$

$$g_t = \sigma\left(W_2\, \mathrm{GELU}\left(W_1 z_t + b_1\right) + b_2\right), \tag{8}$$

where $g_t \in [0, 1]$ denotes the predicted visual dependency of the upcoming token $y_t$. A larger $g_t$ indicates stronger reliance on visual evidence and therefore induces greater visual enhancement. The gate is computed once prior to each Transformer forward pass and shared across all layers.

### 4.4 Visual Dependency Supervision

The gate is designed to predict token-wise visual dependency, requiring supervision that reflects the contextual contribution of visual evidence at each generation step. As shown in Sec. 3.2, visual dependency cannot be reliably determined from static semantic categories. We therefore construct soft supervision labels from the prediction discrepancy between full-vision and vision-drop conditions, enabling fully decoupled training of the gate from the LVLM backbone.

**Offline Label Construction.** We construct supervision labels using teacher-forced inference with the LVLM fully frozen. For each training sample $(I, Q, Y)$, where $Y = \{y_t\}_{t=1}^{T}$ denotes the ground-truth response, we compute the probability of each target token under the full-vision and vision-drop conditions and obtain the visual dependency score $\Delta_t$ defined in Eq. 3. The vision-drop setting retains

visual token positions and embeddings but masks their Key/Value, preserving the input structure and avoiding artifacts introduced by sequence changes. A positive $\Delta_t$ indicates that visual input improves the prediction of $y_t$, while a negative value suggests no positive visual contribution. We truncate negative scores as $\tilde{\Delta}_t = \max(\Delta_t, 0)$ and transform them into soft supervision labels:

$$q_t = 1 - \exp\left(-\tilde{\Delta}_t\right), \qquad q_t \in [0, 1). \tag{9}$$

The resulting $q_t$ serves as the visual dependency label for the generation step $t$. During label construction, we simultaneously cache the corresponding full-vision hidden state $h^L_{t-2}$ and the preceding-token embedding $e(y_{t-1})$, yielding the gate training dataset $\mathcal{D}_G = \{(h^L_{t-2}, e(y_{t-1}), q_t)\}$.

**Decoupled Gate Training.** The gate is trained independently of the LVLM backbone using only the cached features and soft supervision labels, with negligible computational overhead. For each training sample $(h^L_{t-2}, e(y_{t-1}), q_t)$, we construct the gate input $z_t$ and predict the visual dependency $g_t$. The gate is optimized with binary cross-entropy loss:

$$\mathcal{L}_{\text{gate}} = \text{BCE}(g_t, q_t). \tag{10}$$

After training, the gate is integrated into the original decoding pipeline and jointly modulates visual attention with the precomputed visual token importance $s_i$ and layer-wise visual activity $\omega_l$.

# 5 Experiments

## 5.1 Experimental Setup

**Tasks and Benchmarks.** We evaluate ViDRA on three LVLMs: LLaVA-1.5-7B (Liu et al., 2024b), Qwen3-VL-8B (Bai et al., 2025), and InternVL3.5-8B (Wang et al., 2025). Our evaluation considers both visual hallucination mitigation and general vision-language capability. For visual hallucination, we adopt three benchmarks: CHAIR (Rohrbach et al., 2018) for object hallucination in open-ended image captioning, POPE (Li et al., 2023b) for object-existence hallucination, and MMHal-Bench (Sun et al., 2024) for more challenging cases involving attributes, counting, spatial relations, and complex visual reasoning. Following prior work, we randomly sample 500 COCO (Lin et al., 2014) images for CHAIR and POPE. For MMHal-Bench, we follow the standard protocol using 96 image-question pairs across 8 question categories and 12 object topics. For general capability, we adopt VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), VizWiz (Gurari et al., 2018), and TextVQA (Singh et al., 2019) to measure multimodal understanding. More details are provided in Sec. A.1.

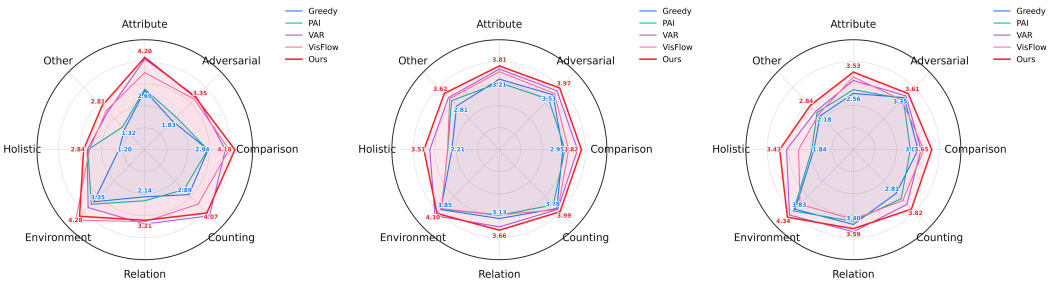
**Implementation Details.** We use a unified hyperparameter setting across all benchmarks without task-specific tuning: $\beta = 1.1$, $\alpha = 0.4$, and $\gamma = 0.7$. All LVLM backbones remain frozen, and all experiments are conducted on a single NVIDIA A6000 GPU.

## 5.2 Main Results

Table 1: Quantitative comparison on visual hallucination benchmarks. Best results are in **bold**, and second-best are underlined.

| Method | LLaVA-1.5-7B | | | | Qwen3-VL-8B | | | | InternVL3.5-8B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $CHAIR_s \downarrow$ | $CHAIR_i \downarrow$ | POPE-Acc ↑ | POPE-F1 ↑ | $CHAIR_s \downarrow$ | $CHAIR_i \downarrow$ | POPE-Acc ↑ | POPE-F1 ↑ | $CHAIR_s \downarrow$ | $CHAIR_i \downarrow$ | POPE-Acc ↑ | POPE-F1 ↑ |
| Greedy | 47.8 | 12.9 | 84.7 | 85.5 | 58.8 | 10.9 | <u>89.4</u> | 89.1 | 39.6 | 9.8 | 86.1 | 87.1 |
| Beam | 48.8 | 13.5 | 85.1 | 84.6 | 55.2 | 10.4 | 89.3 | 89.0 | 36.4 | 9.5 | 86.2 | 87.4 |
| Nucleus | 49.0 | 14.7 | 81.3 | 82.4 | 58.6 | 10.9 | 88.7 | 88.5 | 40.2 | 9.6 | 84.6 | 85.9 |
| VCD | 51.8 | 14.4 | 83.4 | 82.2 | 60.2 | 11.4 | 89.1 | 88.6 | 40.0 | 10.2 | <u>87.2</u> | <u>87.9</u> |
| PAI | <u>34.4</u> | 10.8 | 84.9 | 85.6 | 59.4 | 10.8 | 85.4 | 86.4 | 40.6 | 9.6 | 84.8 | 86.2 |
| VAR | 45.6 | 12.3 | 85.0 | <u>85.9</u> | <u>25.0</u> | <u>8.5</u> | <u>89.4</u> | <u>89.3</u> | <u>33.2</u> | 9.4 | 86.1 | 87.3 |
| VisFlow | 42.1 | <u>9.4</u> | <u>85.5</u> | 85.7 | **24.6** | **8.4** | 88.9 | 88.8 | 34.8 | <u>9.3</u> | 87.0 | 87.7 |
| SADT | 41.8 | 9.9 | 84.2 | 85.1 | 45.8 | 10.0 | 89.1 | 88.7 | 39.4 | 9.7 | 86.0 | 87.1 |
| **Ours** | **31.3** | **8.2** | **87.4** | **88.1** | 25.7 | 8.9 | **89.8** | **90.7** | **31.6** | **9.1** | **88.4** | **89.5** |

We compare ViDRA with existing methods on visual hallucination benchmarks. As shown in Table 1, ViDRA achieves the lowest $CHAIR_s$ and $CHAIR_i$ on LLaVA-1.5-7B and InternVL3.5-8B, outperforming conventional decoding strategies (Greedy, Beam, Nucleus) and existing mitigation methods such as VAR (Kang et al., 2025), VisFlow (Tang et al., 2026), and SADT (Wang et al., 2026b).




(a) LLaVA-1.5-7B (b) Qwen3-VL-8B (c) InternVL3.5-8B

Figure 5: Category-wise performance comparison on MMHal-Bench across three LVLMs.

On POPE, ViDRA achieves the best Accuracy and F1 across all three LVLMs, demonstrating that it effectively reduces object-existence hallucinations without compromising visual discrimination or becoming overly conservative. As illustrated in Fig. 5, on the more challenging and diverse MMHal-Bench, ViDRA outperforms other methods across sub-dimensions such as attribute description, object counting, spatial relations, and holistic understanding, indicating robust mitigation across diverse hallucination types. Overall, ViDRA delivers notable improvements across different LVLMs and evaluation metrics, demonstrating strong cross-model generalization.

Table 2: Quantitative comparison on general vision-language benchmarks.

| Method | LLaVA-1.5-7B | | | | Qwen3-VL-8B | | | | InternVL3.5-8B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | VQAv2 ↑ | GQA ↑ | VizWiz ↑ | TextVQA ↑ | VQAv2 ↑ | GQA ↑ | VizWiz ↑ | TextVQA ↑ | VQAv2 ↑ | GQA ↑ | VizWiz ↑ | TextVQA ↑ |
| Greedy | 78.1 | 61.3 | 54.3 | 48.6 | 83.3 | 61.6 | 69.0 | 83.4 | 80.0 | 61.6 | 60.6 | 78.1 |
| Beam | 77.9 | 61.2 | 54.4 | 48.2 | 83.0 | 61.6 | 68.2 | 82.1 | 77.9 | 60.4 | 59.7 | 76.2 |
| Nucleus | 77.3 | 61.5 | 54.6 | 46.5 | 81.7 | 59.3 | 68.0 | 82.4 | 78.1 | 60.9 | 58.5 | 76.6 |
| VCD | 71.6 | 55.1 | 42.6 | 42.6 | 81.6 | 60.1 | 64.4 | 82.0 | 76.6 | 58.2 | 54.3 | 74.4 |
| PAI | 77.5 | 61.5 | 51.9 | 45.7 | 74.3 | 55.9 | 67.8 | 82.1 | 79.1 | 61.8 | 49.8 | 73.0 |
| VAR | 78.0 | 61.7 | 55.1 | 45.8 | 83.3 | 61.7 | 69.0 | **83.5** | 80.0 | 61.6 | 60.0 | 78.0 |
| VisFlow | 77.6 | 60.9 | 54.1 | 45.7 | 82.2 | 61.7 | 68.7 | 83.0 | 79.9 | 61.5 | 60.1 | 78.2 |
| SADT | 78.2 | 61.9 | 55.3 | 44.1 | 82.6 | 61.2 | 68.6 | 82.3 | 79.8 | 60.7 | 58.2 | 78.0 |
| **Ours** | **78.4** | **62.2** | **56.3** | **48.9** | **84.5** | **62.4** | **69.3** | 83.4 | **81.7** | **63.4** | **60.8** | **78.3** |

Table 2 evaluates whether hallucination mitigation compromises the original generation and reasoning capabilities. Overall, ViDRA consistently preserves or improves performance across all three LVLMs, whereas existing methods often degrade general-task performance. For example, on InternVL3.5-8B, PAI with uniform visual enhancement drops by 10.8 on VizWiz relative to Greedy decoding, suggesting that excessive or indiscriminate visual enhancement may disrupt the model's intrinsic reasoning process. By adaptively modulating visual attention across tokens, layers, and generation steps, ViDRA strengthens visual grounding while improving the general vision-language capability.

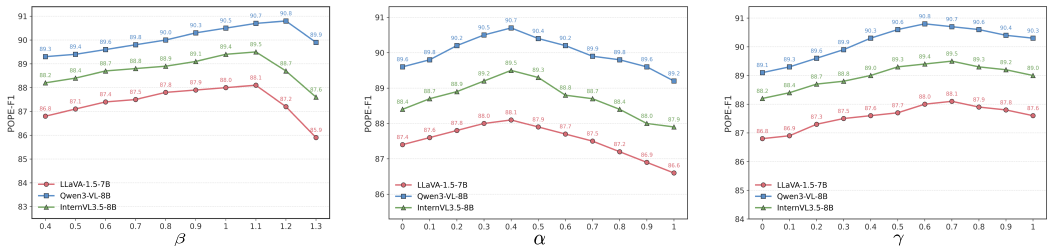
### 5.3 Ablation Studies and Analysis

Table 3: Component ablation of ViDRA on LLaVA-1.5-7B.

| Settings | POPE-Acc ↑ | POPE-F1 ↑ | VQAv2 ↑ | GQA ↑ |
|---|---|---|---|---|
| Full ViDRA | **87.4** | **88.1** | **78.4** | **62.2** |
| w/o Token Importance | 85.9 | 86.8 | 78.0 | 61.4 |
| w/o Layer Activity | 85.8 | 85.6 | 78.2 | 61.8 |
| w/o Dependency Gate | 84.7 | 84.5 | 77.3 | 61.0 |

**Component Ablation.** As shown in Table 3, removing any component consistently degrades both hallucination mitigation and general vision-language performance. The greatest degradation occurs without the Visual Dependency Gate: POPE Accuracy and F1 drop by 2.7 and 3.6, respectively, while VQAv2 and GQA decrease by 1.1 and 1.2. This highlights the importance of modeling step-wise

visual dependency to avoid indiscriminate enhancement. Removing Visual Token Importance or Layer-wise Visual Activity also degrades performance, confirming the benefits of informative token selection and layer-adaptive intervention. Overall, the full ViDRA performs best across all metrics, validating the effectiveness and complementarity of our Token–Layer–Step design.




(a) Global enhancement coefficient $\beta$ (b) Token-importance balancing coefficient $\alpha$ (c) Gate-input fusion weight $\gamma$

Figure 6: Hyperparameter sensitivity analysis of ViDRA across three LVLMs.

**Hyperparameter Analysis.** As shown in Fig. 6, for the global enhancement coefficient $\beta$, ViDRA maintains stable and effective hallucination mitigation within a reasonable range, while overly large values degrade performance by disrupting the original generation balance and producing overly conservative outputs with reduced flexibility and richness. For the token-importance balancing coefficient $\alpha$, $\alpha = 0$ uses only instruction relevance $R_i$, while $\alpha = 1$ relies solely on foreground confidence $P_i$. Instruction relevance alone performs better than foreground confidence alone, while their weighted combination achieves the best performance, confirming their complementarity. For the gate-input fusion weight $\gamma$, $\gamma = 0$ removes the latest generated token from the gate input and yields lower performance. Increasing $\gamma$ improves visual-dependency estimation by incorporating recent semantic information, while an excessively large value weakens the contribution of the preceding multimodal context and degrades performance. Overall, ViDRA remains stable across a broad hyperparameter range, demonstrating strong robustness and validating the default configuration.

Table 4: Inference efficiency comparison on LLaVA-1.5-7B. Latency denotes the average inference time per sample, while TPS measures the number of generated tokens per second.

| Method | Extra Forward | Latency (s) ↓ | TPS ↑ | GPU Memory (GB) ↓ |
|---|---|---|---|---|
| Greedy | No | 1.8 | 38.5 | 13.9 |
| VCD | Yes | 3.5 | 18.8 | 14.6 |
| PAI | Yes | 3.4 | 19.7 | 14.1 |
| VAR | No | 7.5 | 11.4 | 13.6 |
| VisFlow | No | 2.0 | 22.3 | 13.7 |
| SADT | Yes | 3.2 | 17.6 | 16.8 |
| **Ours** | No | 2.2 | 31.5 | 14.2 |

**Inference Efficiency.** As shown in Table 4, ViDRA requires no additional LVLM forward pass and introduces only modest overhead over Greedy decoding, increasing latency from 1.8 to 2.2 s while maintaining 31.5 TPS. Compared with VCD, PAI, VAR, and SADT, our ViDRA achieves higher decoding throughput and substantially lower latency. This efficiency stems from its natural compatibility with the LVLM prefill–decoding architecture: token-level and layer-level priors are computed once during prefill, while decoding only requires a lightweight gate at each step. Overall, ViDRA mitigates visual hallucinations with minimal computational and memory overhead.

## 6 Conclusion

We propose ViDRA, a fine-grained attention rebalancing framework for mitigating visual hallucinations in LVLMs. ViDRA adaptively modulates visual attention: it estimates visual token importance and layer-wise visual activity during prefill, and adjusts enhancement strength via a Visual Dependency Gate during decoding. Experiments show that ViDRA effectively reduces hallucinations while preserving language generation capability and inference efficiency, outperforming existing methods.

## AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and improve the clarity of the manuscript. All AI-assisted content was carefully reviewed and revised by the authors. The research ideas, methodology, experiments, results, and conclusions were independently developed and verified by the authors. We take full responsibility for the final content of this work.

## REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21635–21644. IEEE, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3608–3617, 2018.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27036. IEEE, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023a.

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26700–26710. IEEE, 2024.

Xinlong Li, Di Lin, Shaoyiyi Gao, Yaxuan Liu, Jixian He, Jiaxin Li, Ruonan Liu, Qing Guo, Kairui Yang, and Wei Feng. Hops: Hierarchical open-vocabulary part segmentation with attention-aware filtering and affinity-guided enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27719–27729, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 57689–57733, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024c.

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

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. *arXiv preprint arXiv:2602.22394*, 2026.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9439–9447, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the association for computational linguistics: ACL 2024*, pp. 15840–15853, 2024b.

Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10306–10314, 2026a.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate lvlm object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25315–25325, 2026b.

Ziyang Wang, Mengwei Li, Hao Yin, Wenhao Liu, and Zilei Wang. Posprune: Visual token pruning with positional bias correction for efficient large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10485–10493, 2026c.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10610–10620. IEEE, 2025.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. Mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13040–13051. IEEE, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

# A Experimental Setup and Details

## A.1 Benchmarks and Evaluation Metrics

We evaluate ViDRA on three visual hallucination benchmarks, including CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023b), and MMHal-Bench (Sun et al., 2024), as well as four general vision-language benchmarks, including VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), VizWiz (Gurari et al., 2018), and TextVQA (Singh et al., 2019).

**CHAIR** evaluates object hallucination in open-ended image captioning. We randomly sample 500 images from the COCO (Lin et al., 2014) validation set and evaluate all methods on the same subset for a fair comparison. Given the instruction "Please describe this image in detail.", each model generates a free-form image description. Following the standard CHAIR protocol, generated objects are mapped to the 80 COCO object categories through lexical normalization and synonym matching. An object is considered hallucinated if its corresponding category is absent from the ground-truth object set. We report the instance-level hallucination rate $\text{CHAIR}_i$ and sentence-level hallucination rate $\text{CHAIR}_s$:

$$\text{CHAIR}_i = \frac{N_{\text{hallucinated objects}}}{N_{\text{mentioned objects}}} \times 100\%, \qquad \text{CHAIR}_s = \frac{N_{\text{hallucinated captions}}}{N_{\text{total captions}}} \times 100\%. \tag{11}$$

$\text{CHAIR}_i$ measures the proportion of hallucinated objects among all generated objects, whereas $\text{CHAIR}_s$ measures the proportion of captions containing at least one hallucinated object. Lower values indicate fewer visual hallucinations.

**POPE** evaluates object hallucination through binary visual question answering. Given an image and the question "Is there a <object> in the image?", the model is required to determine whether the queried object actually appears in the image. We evaluate all methods on the same 500 COCO images and identical image-question pairs. POPE includes three negative-sampling settings: Random, Popular, and Adversarial. Random sampling selects absent object categories at random, Popular sampling focuses on frequently occurring object categories, and Adversarial sampling selects absent objects that commonly co-occur with objects present in the image. These settings progressively test whether the model is influenced by dataset frequency and object co-occurrence priors rather than the actual visual content. Positive queries are constructed from objects present in each image, with an even balance of positive and negative samples per evaluation setting. We report Accuracy and F1 score, with higher values indicating better performance.

**MMHal-Bench** assesses hallucination through automated GPT-based evaluation. It consists of 96 image-question pairs that span 12 topics and 8 question categories: attribute, adversarial, comparison, counting, relation, environment, holistic, and others. For each test sample, the evaluator is provided with the image content, question, reference answer, and generated response, and assigns a score from 0 to 6 according to the rubric in Table 5. The scoring criteria jointly consider visual faithfulness and response informativeness. We report the average score across all 96 samples.

**VQAv2** focuses on general visual understanding and reasoning over natural images, for which we evaluate on the test-dev split. **VizWiz** consists of questions posed by blind and low-vision users on images captured under unconstrained real-world conditions, and we use the validation split. **TextVQA** emphasizes reading and reasoning over scene text in natural images and is also evaluated on the validation split. All three benchmarks provide multiple human annotations for each question and adopt the standard VQA accuracy metric. Given a model prediction $p$, the score for each question is calculated as follows:

$$\text{Acc}(p) = \min\left(\frac{N_{\text{match}}(p)}{3}, 1\right), \tag{12}$$

Table 5: Scoring Rubric for MMHal-Bench Evaluation.

| Score | Evaluation criterion |
|---|---|
| 0 | Not informative and contains hallucination |
| 1 | Somewhat informative and contains hallucination |
| 2 | Very informative, but contains hallucination |
| 3 | Not informative, but contains no hallucination |
| 4 | Somewhat informative and contains no hallucination |
| 5 | Very informative and contains no hallucination |
| 6 | Very informative, with good analysis or reasoning, and no hallucination |

where $N_{\text{match}}(p)$ denotes the number of human responses that match the prediction $p$. Predictions matching 3 or more human responses receive a full score of 1. The final accuracy is obtained by averaging the scores over all evaluated questions.

**GQA** evaluates compositional visual reasoning over objects, attributes, and semantic relationships. We use the test-dev-balanced split for evaluation. Each question is associated with a single ground-truth answer, and predictions are evaluated using exact-match accuracy. A prediction is considered correct only if it exactly matches the corresponding ground-truth answer, and the final accuracy is computed as the proportion of correctly answered questions over the entire evaluation set.

### A.2 Baselines and Decoding Configurations

We consider three conventional decoding strategies as reference baselines. **Greedy Decoding** selects the token with the highest conditional probability at each generation step. **Beam Search** (Sutskever et al., 2014) maintains multiple candidate sequences throughout generation and selects the sequence with the highest accumulated score; we set the beam size to 5. **Nucleus Sampling** (Holtzman et al., 2019) introduces stochasticity by sampling from the smallest set of tokens whose cumulative probability exceeds a predefined threshold. We set the top-$p$ to 0.8 and the temperature to 0.7. For open-ended image captioning and visual question answering, where relatively detailed responses are required, we allow a maximum of 512 newly generated tokens. For general vision-language benchmarks that primarily require short-form answers, we follow their standard evaluation settings, setting the maximum number of newly generated tokens to 16 for VQAv2, GQA, and VizWiz, and to 32 for TextVQA.

We further compare ViDRA with five representative hallucination mitigation methods: **VCD**, **PAI**, **VAR**, **VisFlow**, and **SADT**. These baselines cover complementary strategies, including contrastive decoding, attention-based intervention, and visual information regulation. Each method is implemented based on its publicly released code and evaluated using the recommended hyperparameters, without additional tuning. To ensure a fair comparison, all methods use identical backbone checkpoints, prompts, and post-processing procedures.

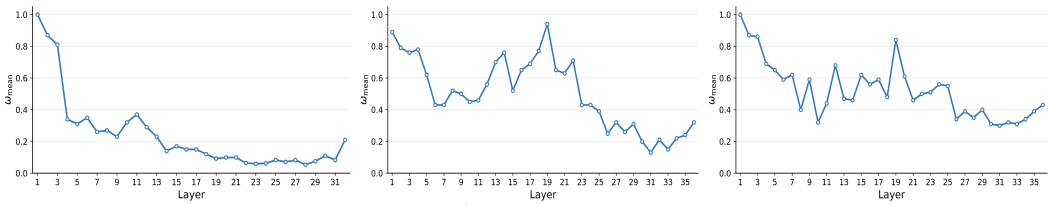
### A.3 Gate Architecture and Training Details

**Architecture.** The Visual Dependency Gate is implemented as a lightweight two-layer MLP with a 256-dimensional bottleneck, following the architecture $4096 \to 256 \to \text{GELU} \to 1$. It contains 1,049,089 trainable parameters (approximately 1.05M).

**Training Setup.** We construct Gate training data from the LLaVA-1.5 instruction-tuning corpus (Liu et al., 2024b), using samples from the Visual Genome (Krishna et al., 2017), GQA (Hudson & Manning, 2019), and OCR-VQA (Mishra et al., 2019). The selected data comprise 260,510 image-text instruction samples and 2,186,156 question-answering turns. We generate token-level visual dependency targets for the response tokens and, after filtering invalid entries, obtain 20,348,979 valid token-level supervision samples. The Gate is optimized using AdamW with a learning rate of $1 \times 10^{-3}$ and a weight decay of $1 \times 10^{-4}$. We apply gradient clipping with a maximum norm of 1.0 and use a batch size of 4,000 token-level samples. Training is performed for a total of 10,000 steps.

# B Additional Experiments and Analysis

## B.1 Analysis of Layer-wise Visual Activity




(a) LLaVA-1.5-7B (b) Qwen3-VL-8B (c) InternVL3.5-8B

Figure 7: Layer-wise visual activity across three LVLMs. Each subplot shows the mean visual activity $\omega_{\text{mean}}$ across 100 sampled images at different Transformer layers.

We analyze the distribution of visual activity across Transformer layers to characterize how different layers engage with visual information. For each layer $l$, we compute the visual activity $\omega_l$ and average it over 100 sampled images. As shown in Fig. 7, visual activity varies substantially across Transformer depths and exhibits distinct layer-wise patterns across LVLMs. In LLaVA-1.5-7B, visual activity is concentrated primarily in the shallow layers and decays rapidly with depth. In contrast, Qwen3-VL-8B and InternVL3.5-8B retain substantial visual activity in intermediate layers, with prominent peaks around the 19th layer. Such cross-model heterogeneity makes a fixed layer-selection strategy difficult to generalize: enhancing only the first few layers may be effective for LLaVA-1.5-7B but would overlook visually active intermediate layers in other LVLMs. These observations motivate our layer-adaptive modulation, which estimates the visual activity of each layer and adjusts the enhancement strength accordingly, thereby avoiding both uniform intervention across layers and model-specific manual layer selection.

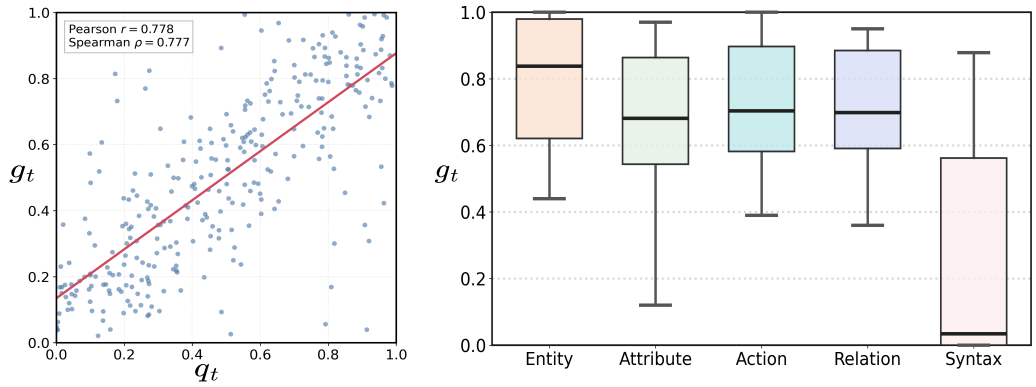
Table 6: Ablation study of ViDRA across different Transformer layer groups and LVLMs. "Greedy" denotes the baseline without visual enhancement. "Shallow", "Middle", and "Deep" apply ViDRA to the first, middle, and last one-third of Transformer layers, respectively, while "All Layers" applies ViDRA to the entire network.

| **Layer Scope** | **LLaVA-1.5-7B** | | **Qwen3-VL-8B** | | **InternVL3.5-8B** | |
|---|---|---|---|---|---|---|
| | POPE-F1 ↑ | VQAv2 ↑ | POPE-F1 ↑ | VQAv2 ↑ | POPE-F1 ↑ | VQAv2 ↑ |
| Greedy | 85.5 | 78.1 | 89.1 | 83.3 | 87.1 | 80.0 |
| Shallow | <u>87.2</u> | <u>78.3</u> | 89.7 | <u>84.2</u> | <u>89.0</u> | <u>81.4</u> |
| Middle | 86.3 | 78.1 | <u>89.8</u> | 84.0 | 88.9 | 81.3 |
| Deep | 85.7 | 78.1 | 89.1 | 83.5 | 87.6 | 80.6 |
| All Layers | **88.1** | **78.4** | **90.7** | **84.5** | **89.5** | **81.7** |

To further investigate the effect of intervention depth, we conduct a layer-selection ablation by dividing the Transformer layers into three groups corresponding to shallow, middle, and deep layers, and restricting ViDRA to each group in turn. As shown in Table 6, the effectiveness of visual intervention varies substantially across both Transformer depth and model. For LLaVA-1.5-7B and InternVL3.5-8B, applying ViDRA to shallow layers yields notably larger gains than restricting it to deeper layers. In contrast, Qwen3-VL-8B benefits similarly from shallow-layer and middle-layer intervention, with the middle-layer setting achieving a slightly higher POPE-F1 score. These differences are consistent with the visual activity patterns observed in Fig. 7, suggesting that a fixed intervention strategy may not generalize well across LVLMs. Applying ViDRA to all layers achieves the best performance, further supporting the effectiveness of our layer-adaptive design.

## B.2 Analysis of Visual Dependency Gate

We further analyze the Visual Dependency Gate to assess whether it captures meaningful token-level variations in visual dependency. We perform two complementary analyses on a separately constructed




(a) Gate Output $g_t$ vs. Dependency Target $q_t$ (b) Gate Output Distribution Across Token Categories

Figure 8: Analysis of the trained Visual Dependency Gate. (a) Predicted gate values exhibit strong agreement with visual dependency targets. (b) Gate responses vary systematically across token categories.

evaluation set that is disjoint from the Gate training data. First, as shown in Fig. 8 (a), the predicted gate output $g_t$ exhibits a strong positive correlation with the corresponding visual dependency target $q_t$, achieving a Pearson correlation coefficient of $0.778$ and a Spearman rank correlation coefficient of $0.777$. This strong agreement indicates that the trained Gate provides a reliable estimate of the visual demand of the upcoming token. Second, as shown in Fig. 8 (b), the gate responses exhibit clear category-dependent patterns that closely align with the visual dependency trends observed in Fig. 3 (a). Together, these results validate the effectiveness of the Visual Dependency Gate.

## C Limitations and Future Work

Despite its effectiveness in mitigating visual hallucinations, ViDRA has several limitations that warrant further investigation.

First, the Visual Dependency Gate relies on a modest amount of offline supervision to learn token-level visual dependency. Although the Gate is lightweight and trained independently of the LVLM backbone, the current framework is therefore not fully training-free. Future work could explore deriving visual dependency directly from pretrained LVLMs.

Second, ViDRA is currently designed primarily for image-based LVLMs. However, its token-level visual dependency modeling and dynamic visual intervention are not inherently limited to static images. Extending the framework to video LVLMs and embodied multimodal models is a promising direction, where visual information evolves over time and dependency patterns may involve richer temporal dynamics and action-related cues.

Finally, the current framework mainly enhances visual information when the predicted visual dependency is positive. In practice, visual signals may also be unnecessary or even distracting for certain generated tokens. A natural extension is therefore to explore bidirectional visual intervention, where visual information is adaptively enhanced or suppressed according to the estimated dependency, enabling finer-grained control over visual information flow during generation.

## D Qualitative Results



**Prompt:** Please describe the image in detail.

**LLaVA:** The image shows a fruit market with a variety of fresh fruits on display. Several pineapples, bananas, and oranges are arranged in baskets, while apples and watermelons can be seen on the counter behind them. A vendor is standing behind fruit stand. The market is crowded with customers.

**LLaVA+PAI:** The image shows many fruits in baskets, and bananas and oranges are in the front. Pineapples are on the left, and there are drinks behind, and more fruits are there on the counter, and the fruits are displayed in the market.

**LLaVA+ViDRA:** The image shows a fruit stand displaying a variety of fresh produce. Bananas and oranges fill the front baskets, with several pineapples arranged on the left. Behind them, colorful drinks and more fruit are displayed on the counter.



**Prompt:** Please describe the image in detail.

**LLaVA:** The image shows a green passenger train traveling along a railway track. Serval passengers can be seen through the windows, while a conductor is standing at the front of the train. The train is passing through a quiet area with houses, trees, and a parking lot nearby.

**LLaVA+PAI:** The image shows a green train that is on the railway track, and the train is in front, and there are tracks beside the train, and there are houses and cars behind the train. The train is near a platform with trees around it.

**LLaVA+ViDRA:** The image shows a green passenger train traveling along a railway track. The train is approaching a station platform, with a second track running alongside it. Trees and houses surround the railway, while several cars are parked near the road.




**Prompt:** Please describe the image in detail.

**LLaVA:** The image shows a red double-decker bus traveling along a city street. The bus is displaying route 15 toward Aldwych, with serval passengers standing on the sidewalk and a driver waving through the front window. A row of red buses is waiting at a nearby bus stop. Tall buildings and trees surround the road.

**LLaVA+PAI:** The image shows a red bus that is a double-decker bus, and the bus is red, and it is driving on the road, and there are people near the bus, and there are buildings behind the bus. The bus has the number 15 and is going toward Aldwych.

**LLaVA+ViDRA:** The image shows a red double-decker bus traveling along a city street. The bus displays route 15 toward Aldwych, while another bus can be seen farther down the road. Pedestrians and trees line the street, with several buildings visible in the background.



**Prompt:** Please describe the image in detail.

**LLaVA:** The image shows several people skiing through a snowy mountain area. The group is preparing their equipment on the snow, with skis, poles, and large backpacks. A ski instructor is giving the group directions while another skier is coming down the slope behind them. A wooden cabin can be seen among the trees in the background.

**LLaVA+PAI:** The image shows several people who are skiing in the snow, and they have skis, and they have ski poles, and they have backpacks, and they are standing in the snow, and there are trees and snow around them. The people are preparing to ski in a snowy area.

**LLaVA+ViDRA:** The image shows a group of people preparing for a ski trip in a snowy mountain area. They are standing on skis and holding ski poles, with several large backpacks and other equipment. Tall evergreen trees cover the surrounding slopes, creating a snowy woodland scene.

Figure 9: Qualitative results on LLaVA-1.5-7B. Red, orange, and green highlight hallucinated content, disfluent expressions, and visually grounded descriptions, respectively.



**Prompt:** Please describe the image in detail.

**Qwen:** The image shows a variety of ingredients and kitchen supplies arranged on a table. There are several bananas, eggs, sugar, olive oil, and other packaged ingredients. A person is preparing a cake beside the table, with a mixing bowl filled with batter. Fresh strawberries and apples are placed next to the bananas.

**Qwen+VAR:** The image shows bananas on a tray, and there are eggs, and there is sugar, and there is oil, and there are some packages, and there is a laptop, and there are things on the table. The ingredients are placed on the table with some kitchen items.

**Qwen+ViDRA:** The image shows a variety of food ingredients and kitchen supplies arranged on a table. Several ripe bananas sit on a tray beside eggs, sugar, and bottles of cooking oil. A laptop and other packaged ingredients are also visible in the background.



**Prompt:** Please describe the image in detail.

**Qwen:** The image shows a young boy sitting on a luggage cart surrounded by several old suitcases and a large wicker basket. His parents are standing nearby, helping him with the luggage. The family is preparing to board a train at the station. The cart is parked outside a building entrance.

**Qwen+VAR:** The image shows a boy who is sitting on the cart, and there are suitcases around him, and there is a basket, and the cart has wheels, and the boy is wearing a yellow shirt, and there are bags behind him. The boy is sitting with the luggage on the cart.

**Qwen+ViDRA:** The image shows a young boy sitting on a red luggage cart surrounded by several vintage suitcases. A large wicker basket and additional luggage are stacked on the cart, while a building entrance is visible in the background.



**Prompt:** Please describe the image in detail.

**Qwen:** The image shows a cozy bedroom with a bed covered by a colorful blanket and several pillows. A teddy bear sits on the headboard, while a small table stands beside the bed. A young girl is sleeping under the blanket, with a book and a glass of water beside her. A window overlooking a garden can be seen on the wall.

**Qwen+VAR:** The image shows a bedroom with a bed, and there is a teddy bear, and there is a small table, and there is a lamp, and there is a blanket, and the walls are purple, and the bed is next to the table. The room has a bed and some furniture in the bedroom.

**Qwen+ViDRA:** The image shows a small bedroom with purple walls and a single bed covered by a colorful patterned blanket. A teddy bear sits on the wooden headboard, while a small round table with a lamp stands beside the bed. A decorative hanging ornament is suspended from the ceiling above the room.



**Prompt:** Please describe the image in detail.

**Qwen:** The image shows a blue and red locomotive traveling through a tropical area at dusk. Tall palm trees surround the railway, while the train's headlight illuminates the tracks ahead. Several passengers can be seen inside the train, waving through the windows.

**Qwen+VAR:** The image shows a train that is blue and red, and the train is on the tracks, and there are palm trees, and the train has a light, and there is grass, and there are trees around the train, and it is at night. The train is moving along the railway near the palm trees.

**Qwen+ViDRA:** The image shows a blue and red locomotive traveling along a railway track at dusk. Tall palm trees surround the track, while the train's headlight illuminates the area in front of it. Grass and vegetation cover the ground around the railway.

Figure 10: Qualitative results on Qwen3-VL-8B. Red, orange, and green highlight hallucinated content, disfluent expressions, and visually grounded descriptions, respectively.



**Prompt:** Please describe the image in detail.

**InternVL:** The image shows a breakfast table with a sandwich, sliced tomatoes, and several drinks. A yellow mug sits beside the plate, while salt and pepper shakers are placed behind the food. A person is sitting at the table and eating the sandwich. A cup of hot coffee is steaming beside the plate.

**InternVL+VisFlow:** The image shows a table with a sandwich, and there are tomatoes, and there is a yellow cup, and there are salt and pepper shakers, and there are glasses, and there are forks and knives, and there is a tablecloth. The sandwich is on a plate and the tomatoes are next to the sandwich.

**InternVL+ViDRA:** The image shows a meal arranged on a patterned tablecloth. A sandwich and several slices of tomato are served on a plate, with a yellow mug and a glass placed beside it. Salt and pepper shakers, utensils, and other tableware are arranged around the meal.



**Prompt:** Please describe the image in detail.

**InternVL:** The image shows two workers loading a large black pipe onto a flatbed truck. The pipe is secured on a yellow support frame, with several containers and equipment nearby. The workers are preparing the pipe for a construction project at a building site. The man on the ground is using a remote

**InternVL+VisFlow:** The image shows a truck with a large pipe on it, and the pipe is black, and there is a man on the truck, and there is another man next to the truck, and the truck is parked, and there are some containers behind the truck. The men are working with the pipe on the truck.

**InternVL+ViDRA:** The image shows two men handling a large black corrugated pipe on a flatbed truck. The pipe is supported by a yellow frame, while several large containers are visible behind the truck. One man stands on the truck bed, while the other stands beside the vehicle.



**Prompt:** Please describe the image in detail.

**InternVL:** The image shows a young boy standing beside a horse and an old wooden cart in a rural area. The boy is holding the horse's reins, while the animal stands beside the cart. The boy is preparing the horse for a tourist ride, and several tourists are waiting nearby. The cart is being used to transport fresh

**InternVL+VisFlow:** The image shows a boy who is standing next to a horse, and the boy is holding the horse, and there is a cart, and the horse is next to the cart, and the boy is wearing a blue shirt, and there are rocks around them, and there are buildings in the background. The boy and the horse are standing near the cart.

**InternVL+ViDRA:** The image shows a young boy standing beside a brown horse attached to an old wooden cart. The boy holds the horse's reins while standing on a rocky dirt road. Low buildings and scattered vegetation can be seen in the background.



**Prompt:** Please describe the image in detail.

**InternVL:** The image shows two firefighters riding in a vintage red fire truck during a parade. The truck is decorated with flags and firefighting equipment, while people watch from the roadside. The firefighters are responding to an emergency at a nearby building, and the driver is sounding the siren.

**InternVL+VisFlow:** The image shows a fire truck that is red, and there are two firefighters, and they are sitting on the fire truck, and there are flags on the truck, and there are firefighting tools on the truck, and the truck is on the road, and there are trees behind the truck. The firefighters are riding in the fire truck.

**InternVL+ViDRA:** The image shows two firefighters riding in a vintage red fire truck on a city street. The truck is decorated with flags and equipped with firefighting tools, while another person can be seen standing behind the firefighters. Leafless trees and buildings line the street in the background.

Figure 11: Qualitative results on InternVL3.5-8B. Red, orange, and green highlight hallucinated content, disfluent expressions, and visually grounded descriptions, respectively.