

# DICA: Dual-Indicator Guided Contrastive Alignment in Multimodal Large Language Models

Hao Yang, Jin Wang\* and Xuejie Zhang  
School of Information Science and Engineering  
Yunnan University  
Kunming, China

yanghao888@stu.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

## Abstract

Human visual reasoning typically follows a coarse-to-fine attention process, starting from global scene understanding and gradually focusing on question-relevant regions. However, multimodal large language models may deviate from this pattern due to attention drift and the underutilization of visual evidence, which can lead to hallucinations. To mitigate these issues, this study proposes a Dual-Indicator Guided Contrastive Alignment (DICA), which tracks two information-theoretic indicators during inference: Visual Attention Entropy (VAE), which reflects the concentration of visual attention, and Output Image Correlation (OIC), which measures the dependence of generated outputs on the visual input. An abnormal increase in VAE or a decrease in OIC corresponds to different failure modes, which trigger targeted contrastive alignment to restore visual grounding. Experimental results across multiple benchmarks demonstrate that DICA consistently outperforms existing approaches and substantially reduces hallucinations, highlighting the effectiveness of indicator-driven intervention in improving multimodal inference reliability. The code is publicly available at <https://github.com/BGWH123/DICA/>.

## 1 Introduction

For Visual Question Answering (VQA), human visual attention follows a coarse-to-fine paradigm, typically begin by conducting a global scan of the scene to acquire an overall semantic structure and contextual cues. As the question is progressively interpreted, attention gradually narrows and shifts toward the local regions most relevant to the query, and the final answer becomes highly dependent on these key visual details (Navon, 1977). This transition from global context awareness to fine-grained semantic focus has been repeatedly vali-

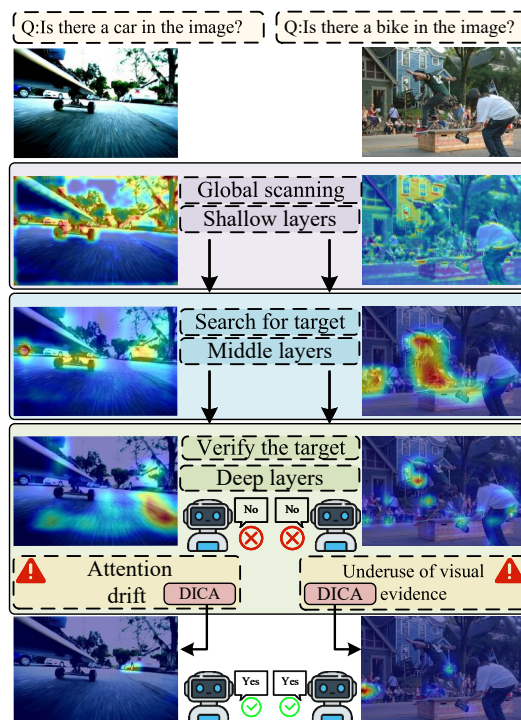


Figure 1: An example of attention drift and underuse of visual evidence.

dated in cognitive and vision studies (Das et al., 2017; Yarbus, 2013; Itti and Koch, 2001).

To investigate whether modern Multimodal Large Language Models (MLLMs) exhibit a similar attention pattern, a systematic empirical analysis is conducted in Section 2. The results reveal a consistent transition from global attention to localized focus. As reasoning unfolds, the MLLMs’ visual attention progressively shifts from a diffuse global distribution to a concentrated focus on regions semantically aligned with the query.

Besides, two major issues are found in the experiments, as shown in Figure 1. **(a) Attention Drift:** During reasoning, the model’s visual attention may suddenly disperse or shift to regions that are semantically irrelevant to the query. **(b) Underuse of Visual Evidence:** The model’s reliance on visual

\*Corresponding author.

cues does not consistently escalate as the network deepens. Conversely, in some cases, the reliance on visual content even degrades in deeper layers. These two issues consequently cause MLLMs to yield incorrect response.

However, existing works, including both training-based approaches (Cai et al., 2024; Villa et al., 2025; Liu et al., 2023a; Yu et al., 2024) and training-free inference-time methods (Leng et al., 2024; Huang et al., 2024; Favero et al., 2024; Woo et al., 2025; Chen et al., 2025), generally treat hallucination as a monolithic phenomenon and lack explicit mechanisms to diagnose its root causes. They fail to distinguish whether hallucinations arise from attention drift or from underuse of visual evidence, and therefore apply uniform correction strategies that may be misaligned with the actual failure modes. As a result, these methods are often unable to perform precise, adaptive interventions when different hallucination patterns occur during reasoning.

In contrast, we propose Dual-Indicator Guided Contrastive Alignment (DICA), which explicitly diagnoses and addresses different hallucination sources during inference. Specifically, DICA continuously monitors two typical failure patterns, attention drift and underuse of visual evidence, via dedicated indicators. Once either indicator exceeds a predefined threshold, a targeted correction is triggered. If attention drift is detected, the hidden states of a neutral query are incorporated to calibrate the current hidden states; if underuse of visual evidence is detected, the original prediction is calibrated using the best alternative distribution.

Extensive experiments on multiple benchmarks show that DICA consistently enhances reasoning accuracy, validating both our theoretical analysis and the practical efficacy of the proposed approach.

The contributions are summarized as follows:

- We empirically uncover two key bottlenecks about MLLM hallucinations: **(a) attention drift** and **(b) underuse of visual evidence**.
- DICA is proposed as a method that can be applied to different models and performs distribution correction at inference time, without requiring retraining or changes to the model architecture.
- Extensive evaluations demonstrate that DICA significantly boosts reasoning accuracy and suppresses hallucinations by effectively

strengthening the model’s reliance on visual evidence.

## 2 Motivation

### 2.1 Indicator Definition

To analyze whether MLLMs exhibit attention patterns similar to human visual attention and whether their answers remain grounded in visual cues, we introduce two diagnostic metrics<sup>1</sup>: **(1) Visual Attention Entropy (VAE)**, which captures the concentration level of visual attention, and **(2) Output–Image Correlation (OIC)**, which measures the model’s reliance on visual cues. These metrics serve as diagnostic indicators for identifying abnormal attention and evidence-utilization behaviors rather than causal explanations of hallucination.

**Visual Attention Entropy.** Shannon entropy (Shannon, 1948) is employed to quantify the concentration of attention over visual tokens. Let the attention weights over all tokens be  $\{a_i^{(l,t)}\}_{i=1}^T$  at layer  $l$  and decoding step  $t$ , and let  $m \in \{0, 1\}^T$  denote the image token mask, where  $m_i = 1$  indicates a visual token. The attention weights corresponding to visual tokens are first selected:

$$\mathcal{A}^{(l,t)} = \{a_i^{(l,t)} \mid m_i = 1\}. \quad (1)$$

These are then normalized to form a probability distribution:

$$p_i^{(l,t)} = \frac{a_i^{(l,t)}}{\sum_{j:m_j=1} a_j^{(l,t)}}, \quad i \in \{j \mid m_j = 1\}. \quad (2)$$

The VAE for layer  $l$  at step  $t$  is computed as

$$\text{VAE}^{(l,t)} = - \sum_{i:m_i=1} p_i^{(l,t)} \log p_i^{(l,t)}. \quad (3)$$

The sequence  $\{\text{VAE}^{(l,1)}, \dots, \text{VAE}^{(l,T)}\}$  represents the evolution of the model’s visual attention across decoding steps for each layer and token. Higher entropy indicates more dispersed attention, while lower entropy indicates more concentrated attention over a subset of visual tokens.

**Output–Image Correlation.** OIC measures the influence of image-conditioned hidden states on the model’s output distributions relative to unconditioned hidden states.

Let  $h^{(l,t)}$  denote the hidden states at decoding step  $t$  and layer  $l$  when conditioned on both text and

<sup>1</sup>The theoretical basis and mathematical proofs for these metrics are provided in Appendices A and B, respectively.

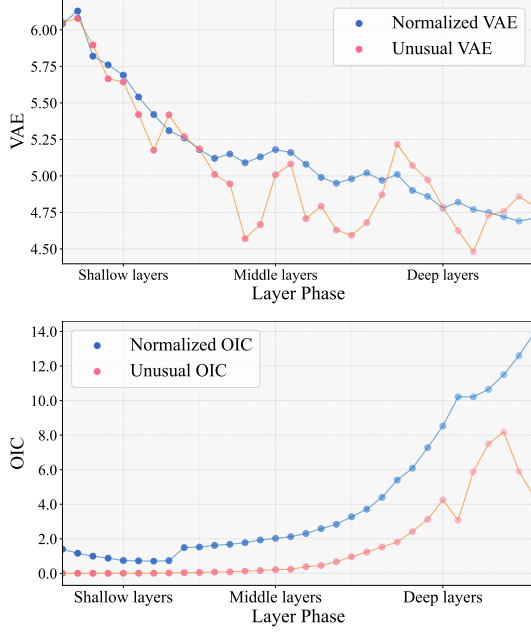


Figure 2: Top: Evolution of VAE across layer phases. Bottom: Evolution of OIC across layer phases.

image, and let  $h_{wi}^{(l,t)}$  denote the hidden states when conditioned only on text, without image. The corresponding output logits are obtained via a linear projection:

$$\mathbf{z}^{(l,t)} = \mathbf{h}^{(l,t)} W^\top, \quad (4)$$

$$\mathbf{z}_{wi}^{(l,t)} = \mathbf{h}_{wi}^{(l,t)} W^\top, \quad (5)$$

where  $W$  is the language modeling head parameters matrix. The probability distributions over the vocabulary are then computed using the softmax function:

$$\mathbf{p}^{(l,t)} = \text{softmax}(\mathbf{z}^{(l,t)}), \quad (6)$$

$$\mathbf{p}_{wi}^{(l,t)} = \text{softmax}(\mathbf{z}_{wi}^{(l,t)}). \quad (7)$$

The OIC for each token is defined:

$$\text{OIC}^{(l,t)} = \sum_v p_v^{(l,t)} \left( \log p_v^{(l,t)} - \log p_{wi,v}^{(l,t)} \right), \quad (8)$$

where  $v$  indexes the vocabulary. The sequence  $\{\text{OIC}^{(l,1)}, \text{OIC}^{(l,2)}, \dots, \text{OIC}^{(l,T)}\}$  then represents the evolution of the model’s output dependence on image information across decoding steps for each layer and token. Higher OIC values indicate stronger reliance on visual input, while lower values indicate that the output primarily depends on text context.

## 2.2 Findings

### Insight 1

Visual attention exhibits a transition from diffuse to more concentrated patterns, as reflected by a generally decreasing VAE.

We systematically study cross-modal attention in LLaVA-1.5 (Liu et al., 2024b) across transformer layers. Specifically, we isolate the attention tokens aligned with visual tokens and calculate their VAE to measure the degree of focus or dispersion in the attention distribution. As shown at the top of Figure 2, it is observed that the downward trajectory of VAE largely supports hypothesis of a global-to-local attention evolution.

However, during decoding, VAE occasionally shows sharp increases. These peaks indicate sudden rises in visual attention entropy, meaning the model’s attention becomes unexpectedly dispersed. Such behavior typically signals confusion or a disconnect from visual evidence and may foreshadow upcoming hallucinated content, suggesting that focused visual cues no longer support the generated output. This anomaly will be further discussed in 2.3.

### Insight 2

During the reasoning process, the model progressively increases its reliance on visual information, as evidenced by rising OIC values.

The model’s reliance on visual information was further quantified by the OIC metric, as illustrated in the bottom part of Figure 2. Under normal reasoning, OIC values typically show a gradual increase or remain high, indicating that the model can consistently and effectively leverage cross-modal mappings to incorporate image features when constructing its answers, thereby ensuring that the generated text is grounded in visual evidence.

However, at certain moments, the metric exhibits abnormal sudden drops. Such abrupt decreases signal a sharp reduction in the model’s dependence on visual input, with the generation distribution becoming primarily driven by the internal priors of the language model. This sudden disengagement from the visual context often implies that the model is performing blind inference based solely on its textual priors, which can easily lead to hallucinated outputs that are inconsistent with the image content. This anomaly will be discussed in Section 2.3.

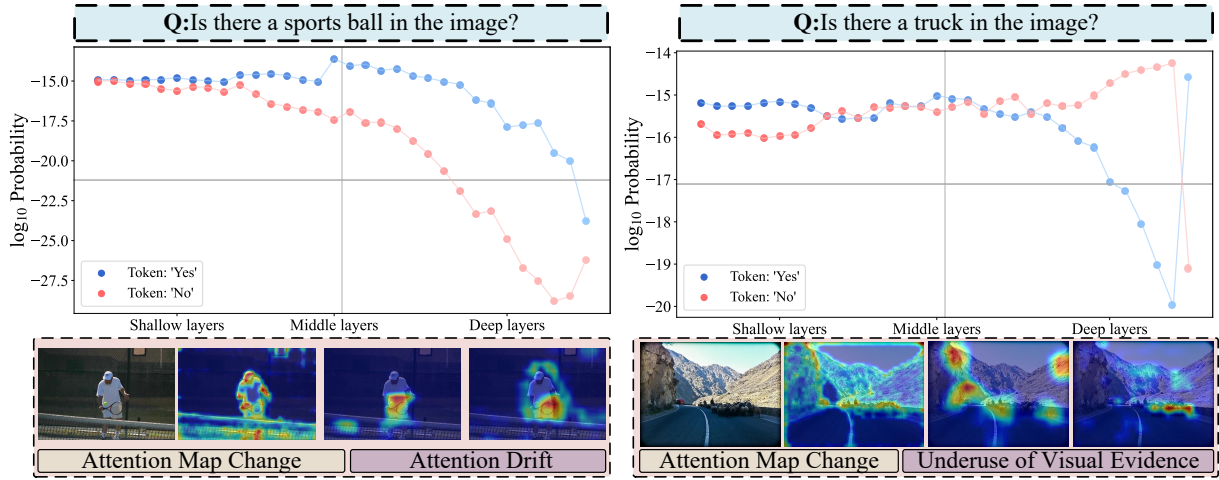


Figure 3: Evolution of answer log-probabilities and attention maps across layers.

### 2.3 Anomaly Analysis

In the preceding analysis of indicators, two types of abnormal phenomena have been identified: **sudden increases in VAE and drops in OIC**. To further investigate the consequences of these two situations, a step-by-step visualization of the reasoning trajectories in which such abnormalities occur is conducted. This includes examining layer-wise changes in attention heatmaps as well as the evolution of the answer log-probabilities, as illustrated in Figure 3. After correction by DICA, the results are provided in D.3.5.

When a sudden surge in VAE is observed, the visual attention previously concentrated on relevant regions abruptly disperses or drifts toward areas semantically irrelevant to the question. Such attention drift interrupts the expected global-to-local focusing process and prevents the model from maintaining stable reliance on critical visual evidence in subsequent layers. This phenomenon is referred to as **attention drift**. The corresponding answer probability curves also exhibit pronounced fluctuations, indicating that the model progressively deviates from a visually grounded reasoning trajectory and becomes prone to erroneous or hallucinatory predictions. The left of Figure 3 shows a representative example of this phenomenon.

In contrast, when OIC experiences a decline, the output distribution reflects a clear reduction in reliance on image information. In this situation, the generation process becomes increasingly dominated by language priors, causing the model to ignore visual evidence and shift toward prior-driven inference, i.e., **underuse of visual evidence**. During this phase, the answer become unstable and

decline, eventually diverging from the actual visual content. As shown in the right of Figure 3, this lack of visual grounding similarly leads to incorrect predictions.

### 3 Dual-Indicator Guided Alignment

Based on the above analysis, the rise in VAE and the drop in OIC lead to attention drift and loss of visual grounding, which in turn cause reasoning errors. Thus, a suitable method should address these issues and maintain the model’s alignment with visual evidence. Therefore, the DICA architecture is proposed, as shown in Figure 4. It is composed of two components, namely VAE-Guided Drift Correction and OIC-Guided Grounding Restoration.

At each MLLM layer  $l$  and decoding step  $t$ , the two indicators are computed:

$$\text{VAE}^{(l,t)} = \text{VAE}(a^{(l,t)}), \quad (9)$$

$$\text{OIC}^{(l,t)} = \text{OIC}(\mathbf{h}^{(l,t)}, \mathbf{h}_{\text{wi}}^{(l,t)}), \quad (10)$$

where both functions follow the definitions provided in Section 2.1.

#### 3.1 VAE-Guided Drift Correction

Attention drift is identified when the VAE surpasses a predefined threshold corresponding to a proportion  $\alpha_{\text{VAE}}$  of its dynamic range, indicating dispersion of visual attention at the feature level. In this state, the hidden representations of image tokens in the Normal path are contaminated by high-entropy noise, causing the model to disperse focus toward irrelevant regions.

To correct this feature-level deviation, we leverage the Descriptive path. Due to its inherently descriptive generation process, the Descriptive path

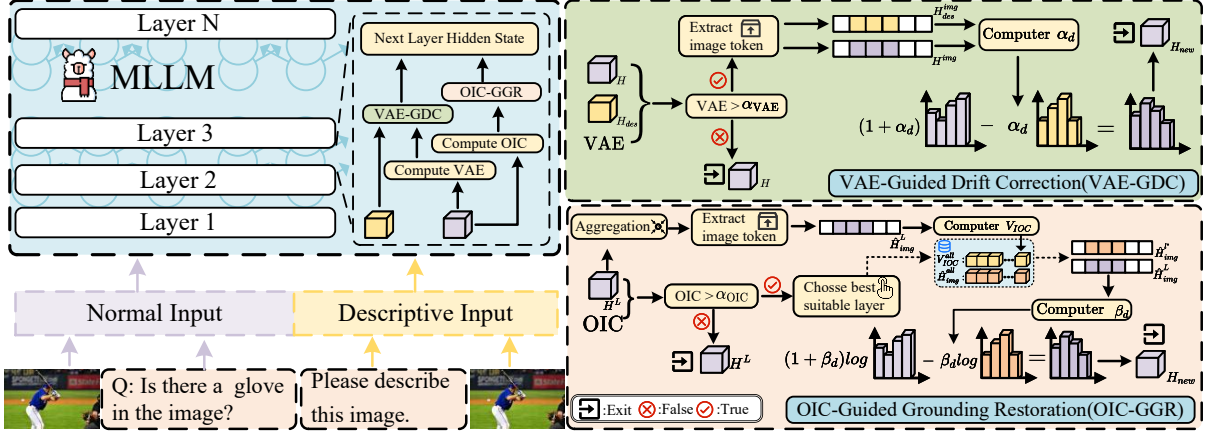


Figure 4: The overall architecture of DICA.

captures richer fine-grained details (Li et al., 2023a; Peng et al., 2023; Liu et al., 2023b) and maintains a more uniform attention distribution across regions. Based on this property, we perform contrastive decoding directly on hidden states to mitigate the effect of high-entropy contamination.

Specifically, let  $H^{\text{img}}$  and  $H_{\text{desc}}^{\text{img}}$  denote the image-token hidden states in the Normal and Descriptive paths, respectively. We map them to logits  $Z$  and  $Z_{\text{desc}}$  via the language modeling projection layer. Then, a dynamic weight is computed  $\alpha_d$  to control the correction strength based on their distributional discrepancy:

$$\alpha_d = \text{clip} \left( 1 - \text{JSD}(Z \| Z_{\text{desc}}), 0, 1 \right), \quad (11)$$

where  $\text{JSD}(\cdot)$  measures the Jensen–Shannon Divergence (Lin, 1991) between the hidden-state feature distributions. The corrected hidden state is then computed as:

$$H^{\text{new}} = (1 + \alpha_d) H^{\text{img}} - \alpha_d H_{\text{desc}}^{\text{img}}. \quad (12)$$

### 3.2 OIC-Guided Grounding Restoration

For step  $t$  of a given sample  $b$ , if the OIC value shows a persistent decrease and falls below a threshold defined as a proportion  $\alpha_{\text{OIC}}$  of its own range, we consider this an indication of underuse of visual evidence.

After computing OIC, we further require a metric that explicitly captures whether the model attends to image regions during reasoning. To this end, we define the image token attention vector  $V_{\text{OIC}}$ , which measures the explicit attention strength that the current language token allocates to the image tokens.

Formally, let  $\mathbf{A} \in \mathbb{R}^{B \times H \times T \times T}$  denote the multi-head attention scores of a given layer  $l$ , where  $B$  is the batch size,  $H$  is the number of attention heads, and  $T$  is the length of the input sequence. Let  $\mathbf{m} \in \mathbb{R}^{B \times T}$  be a binary mask indicating the positions of the image tokens within the sequence. At the decoding step  $t$ , the model uses the last query position to attend to all key positions. Following prior findings (Neo et al., 2024; Xu et al., 2025) that the maximum activation across heads provides a reliable estimate of the model’s dominant focus,  $\mathbf{A}$  is aggregated across the head dimension as:

$$\hat{\mathbf{A}}_{b,t,j} = \max_{h \in \{1, \dots, H\}} \mathbf{A}_{b,h,t,j}, \quad \hat{\mathbf{A}} \in \mathbb{R}^{B \times T \times T}, \quad (13)$$

where  $j$  indexes the source positions attended to.

The aggregated attention distribution is then restricted to the positions corresponding to image tokens, as specified by the mask  $\mathbf{m}$ . For each sample  $b$ , decoding step  $t$  and layer  $l$ , the raw image-directed attention vector is

$$\mathbf{v}_{b,t} = \left[ \hat{\mathbf{A}}_{b,t,j} \mid \mathbf{m}_{b,j} = 1 \right]. \quad (14)$$

To obtain a normalized attention profile over image tokens, we apply a softmax operation:

$$V_{\text{OIC}}^{(b,t)} = \text{softmax}(\mathbf{v}_{b,t}), \quad (15)$$

where  $V_{\text{OIC}}^{(b,t)}$  represents the relative attention probability that the model allocates to the image token at the current decoding step in the  $l$ -th layer. For simplicity, we denote it as  $V_{\text{OIC}}^l$ .

The JSD is then computed between this layer and each preceding layer to quantify the layer-wise difference in visual attention distributions.

$$\text{dist}(V_{\text{OIC}}^N, V_{\text{OIC}}^L) = \text{JSD}(V_{\text{OIC}}^N \| V_{\text{OIC}}^L), \quad (16)$$

where  $L$  denotes the final decoding layer. The layer  $N = 1, \dots, L - 1$ . Exhibiting the maximal divergence is selected as the target intermediate layer  $l^*$ :

$$l^* = \arg \max_{N=1, \dots, L-1} \text{dist}(V_{\text{OIC}}^{(N)}, V_{\text{OIC}}^{(L)}). \quad (17)$$

To adaptively modulate the strength of grounding restoration, we define a dynamic scaling factor  $\beta_t$  based on the inter-layer divergence of visual attention distributions:

$$\beta_d = \text{clip} \left( \text{JSD}(V_{\text{OIC}}^{l^*} \parallel V_{\text{OIC}}^L), 0, 1 \right), \quad (18)$$

Once the layer  $l^*$  is selected, intermediate-layer contrastive decoding is performed by comparing the predicted token distributions from the final layer and the selected intermediate layer. Let  $H_{\text{img}}^L$  and  $H_{\text{img}}^{l^*}$  denote the vocabulary logits produced by the final and intermediate layers under image-aware hidden representations. The contrastive decoding distribution is then computed as:

$$H_{\text{new}} = \text{softmax} \left( (1 + \beta_d) \log H_{\text{img}}^L - \beta_d \log H_{\text{img}}^{l^*} \right). \quad (19)$$

## 4 Experiments

### 4.1 Experiments Setup

**Datasets and Metrics.** We evaluate DICA on three benchmarks as follows:

**(1) POPE (Li et al., 2023d).** The benchmark is designed to test whether MLLMs can correctly judge the presence or absence of specific objects in an image, thereby identifying object-level hallucinations. It formulates the task as Yes/No queries based on object annotations. The dataset is balanced (50% real objects and 50% non-existent objects) and employs three sampling strategies: random, popular, and adversarial. Built from MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019), POPE contains approximately 27k query-answer pairs for evaluation.

**(2) MME (Zhang et al., 2021).** The comprehensive benchmark is designed to assess MLLM’s performance across multiple dimensions. It covers ten perception-oriented tasks and four cognition-related tasks. Object-level hallucination is examined through subsets evaluating object presence and counting, while attribute-level hallucinations are analyzed via subsets focusing on object color and position.

**(3) CHAIR (Rohrbach et al., 2018).** In this evaluation framework, object hallucination in image captioning is assessed by prompting MLLMs to produce descriptive captions for a randomly chosen subset of 500 images from the MS-COCO validation set. The metric quantifies hallucination by measuring the fraction of objects referenced in the generated caption that are absent from the corresponding ground-truth annotations. It provides two primary metrics:  $\text{CHAIR}_i$  at the object-instance level and  $\text{CHAIR}_s$  at the sentence level.

**Implementation Details.** In experiments, we selected the models LLaVA-1.5 (Liu et al., 2024b) and Qwen-VL (Bai et al., 2023) as the backbones. The hyperparameters are set as  $\alpha_{\text{VAE}} = 0.8$ , and  $\alpha_{\text{OIC}} = 0.5$ .<sup>2</sup> All experiments were conducted on an NVIDIA A100 GPU.

**Baselines.** All baseline methods in our experiments are selected from VCD (Leng et al., 2024), OPERA (Huang et al., 2024), M3iD (Favero et al., 2024), AvisC (Woo et al., 2025), and MoD (Chen et al., 2025). These methods aim to mitigate hallucinations by reducing reliance on language priors during the decoding process.

### 4.2 Main Results

**Results on POPE.** Table 1 compares DICA with representative baselines on the POPE benchmark. DICA achieves the best performance in nearly all evaluation settings and datasets. On LLaVA-v1.5 under the AOKVQA/Popular setting, DICA reaches an F1-score of 85.99, outperforming all baselines. The concurrent improvements in accuracy and F1-score indicate that DICA effectively reduces false positive predictions, making the model more conservative in its production of affirmative answers. This directly alleviates object hallucination and improves fine-grained object understanding.

**Results on CHAIR.** Table 2 shows that DICA substantially reduces hallucinations in generative tasks. On both LLaVA-v1.5 and Qwen-VL, DICA achieves the lowest sentence-level and instance-level CHAIR scores among all methods. On LLaVA-v1.5, DICA reduces  $\text{CHAIR}_s$  to 41.8 and  $\text{CHAIR}_i$  to 11.5 while maintaining high recall, indicating that hallucination suppression does not come at the cost of descriptive coverage. On Qwen-VL, DICA attains similarly low hallucination rates with short outputs and preserves reasonable recall.

<sup>2</sup>Hyperparameter analysis is provided in D.3.2.

Setting	Method	MSCOCO				AOKVQA				GQA			
		LLaVA-v1.5		Qwen-VL		LLaVA-v1.5		Qwen-VL		LLaVA-v1.5		Qwen-VL	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Random	sample	83.80	84.20	84.90	82.90	81.80	83.50	86.80	85.80	81.60	83.50	83.30	79.20
	AvisC	82.30	83.50	82.90	80.00	79.10	82.10	84.70	83.00	79.00	82.20	80.50	77.80
	VCD	85.00	84.20	85.50	83.60	81.20	83.20	87.40	86.60	82.20	84.10	82.00	80.50
	OPERA	89.20	88.81	<u>87.31</u>	<b>86.92</b>	<u>88.02</u>	84.59	<u>88.19</u>	<u>88.43</u>	<u>88.13</u>	<b>88.91</b>	86.02	85.29
	M3ID	86.10	85.00	85.30	83.40	82.90	84.60	87.10	85.90	83.30	84.50	82.40	79.70
	MoD	89.20	<b>89.10</b>	86.00	84.10	86.50	87.50	87.80	86.90	86.20	87.40	83.80	82.30
	DICA (ours)	<b>89.23</b>	<u>88.82</u>	<b>87.37</b>	<u>85.97</u>	<b>89.23</b>	<b>89.42</b>	<b>89.47</b>	<b>88.93</b>	<b>89.09</b>	<u>88.65</u>	<b>86.67</b>	<b>85.42</b>
	w/o VAE-GDC	86.50	85.04	86.01	84.19	88.60	87.78	87.23	86.48	88.40	88.06	83.57	81.93
	w/o OIC-GGR	87.76	86.67	86.13	84.24	88.90	89.07	89.15	88.58	88.57	88.11	86.23	84.77
Popular	sample	82.00	82.60	84.00	81.90	75.30	78.70	85.60	84.70	73.10	77.50	75.90	74.90
	AvisC	78.20	80.50	82.80	80.10	71.80	77.20	83.90	83.40	67.40	74.80	74.20	72.30
	VCD	82.10	83.20	84.90	83.60	74.70	78.50	86.30	85.10	71.50	76.80	75.90	75.60
	OPERA	86.64	<u>86.62</u>	87.44	<b>86.68</b>	<u>83.22</u>	<u>84.67</u>	87.91	<u>87.13</u>	<u>79.27</u>	<u>82.11</u>	81.97	<b>82.12</b>
	M3ID	82.80	84.10	84.20	82.70	75.80	79.40	85.90	84.60	72.30	77.10	76.80	77.00
	MoD	85.70	86.10	85.60	83.70	79.50	82.20	86.50	85.60	74.00	78.60	79.80	78.90
	DICA (ours)	<b>87.72</b>	<b>86.65</b>	<b>87.50</b>	<u>86.13</u>	<b>86.64</b>	<b>85.99</b>	<b>88.19</b>	<b>87.50</b>	<b>81.88</b>	<b>82.52</b>	<b>82.13</b>	<u>81.72</u>
	w/o VAE-GDC	86.40	84.92	85.50	83.57	85.78	85.87	86.97	86.10	79.43	81.08	78.93	77.90
	w/o OIC-GGR	86.53	85.06	85.63	83.75	86.46	85.83	87.03	86.19	81.26	82.02	79.13	78.07
Adversarial	sample	75.80	78.10	82.10	80.20	67.40	73.70	80.40	80.50	68.00	74.50	75.50	74.40
	AvisC	74.20	77.70	81.20	78.50	64.40	73.00	78.10	77.30	64.10	72.90	75.50	73.40
	VCD	76.30	78.70	84.00	82.00	68.10	74.60	80.70	80.80	67.60	74.50	76.70	76.20
	OPERA	81.24	81.38	84.78	83.45	<u>73.82</u>	<u>77.91</u>	80.82	<b>81.54</b>	75.00	78.71	<b>80.24</b>	<b>80.64</b>
	M3ID	77.10	78.90	83.20	80.80	68.30	74.60	80.50	80.40	67.20	74.00	77.10	76.70
	MoD	79.70	81.30	84.00	82.30	69.10	75.40	81.00	80.90	68.70	75.30	78.90	78.10
	DICA (ours)	<b>83.20</b>	<b>82.55</b>	<b>84.80</b>	<b>83.67</b>	<u>77.77</u>	<b>78.63</b>	<b>81.33</b>	<u>80.98</u>	<b>78.40</b>	<b>79.86</b>	<u>79.23</u>	<u>78.21</u>
	w/o VAE-GDC	81.40	81.70	83.80	81.98	76.47	76.41	81.07	80.68	77.79	79.37	78.57	78.53
	w/o OIC-GGR	82.37	81.94	84.00	82.24	77.47	76.45	81.23	81.16	77.53	80.41	78.80	77.89

Table 1: Results on POPE benchmark. *Acc* and *F1* denote Accuracy and F1 score, respectively. **Bold** indicates the best performance, and underline denotes the second best.

DICA exhibits consistent hallucination suppression across architectures, demonstrating strong robustness.

**Results on MME.** As reported in Table 3, DICA achieves the highest overall scores across all settings. On LLaVA-v1.5, DICA attains a total score of 661.5, exceeding all baseline methods. It performs best on the all subset and maintains leading performance on Counting and Position, indicating reliable perception of object presence and attributes. On Qwen-VL, DICA again achieves the highest total score of 648.3, with clear gains on the Existence, Position, and Color subsets. This suggests that DICA benefits from stronger visual encoders and improves cross-modal alignment. While performance on the Count subset is slightly lower than the best baseline, DICA remains the most stable method overall.

### 4.3 Ablation Studies

To verify the effectiveness of the two key components in DICA, we conduct ablation studies on the POPE benchmark. As shown in Table 1, removing either module results in a noticeable performance

	Method	CHAIR <sub>s</sub> ↓	CHAIR <sub>i</sub> ↓	Recall ↑	Length
LLaVA-v1.5	sampling	52.8	15.9	77.3	93.4
	VCD	51.0	14.9	77.2	101.9
	M3iD	56.2	17.0	79.3	97.1
	AvisC	44.0	13.7	72.9	89.8
	MoD	<u>42.6</u>	<u>12.4</u>	78.9	97.6
	DICA (ours)	<b>41.8</b>	<b>11.5</b>	77.8	95.8
Qwen-VL	sampling	2.8	3.0	31.0	5.3
	VCD	1.4	1.2	30.8	4.0
	M3iD	1.7	1.3	31.8	3.4
	AvisC	1.6	1.6	32.0	4.4
	MoD	<u>0.8</u>	<u>1.0</u>	32.1	3.8
	DICA (ours)	<b>0.8</b>	<b>1.0</b>	27.0	5.7

Table 2: Results on CHAIR benchmark. Lower CHAIR<sub>s</sub> and CHAIR<sub>i</sub>, along with higher Recall, correspond to better performance.

drop, confirming that both VAE-GDC and OIC-GGR are indispensable. The degradation is more pronounced when VAE-GDC is removed. This is primarily because attention drift occurs more frequently during inference, as indicated by the trigger statistics in Figure 6. Moreover, VAE-GDC operates at the attention level, directly reinforcing the visual grounding foundation, while OIC-GGR intervenes later at the output distribution level. Consequently, VAE-GDC serves a more foundational role

Method	Object-level		Attribute-level		Total	
	Existence	Count	Position	Color		
LLaVA-v1.5	sampling	170.0	103.3	108.3	128.3	510.0
	VCD	180.0	110.0	108.3	133.3	531.7
	M3ID	185.0	118.3	121.7	128.3	553.3
	AvisC	195.0	116.7	131.7	153.3	596.7
	MoD	195.0	141.7	126.7	175.0	638.3
	DICA (ours)	<b>195.0</b>	<b>148.3</b>	<b>143.2</b>	<b>175.0</b>	<b>661.5</b>
Qwen-VL	sampling	160.0	143.3	113.3	165.0	581.7
	VCD	165.0	140.0	113.3	175.0	593.3
	M3ID	165.0	143.3	103.3	175.0	586.7
	AvisC	160.0	145.0	113.3	160.0	578.3
	MoD	170.0	<b>160.0</b>	103.3	180.0	613.3
	DICA (ours)	<b>170.0</b>	155.0	<b>133.3</b>	<b>190.0</b>	<b>648.3</b>

Table 3: Results on MME benchmark. The performance is measured by MME Score. The *Total* column represents the sum of four individual results in each row.

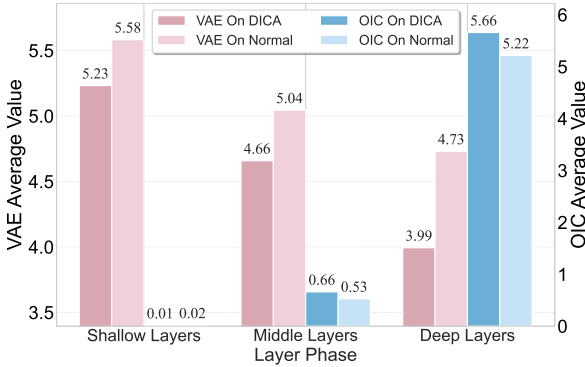


Figure 5: The impact of DICA on VAE and OIC within POPE.

in suppressing hallucinations. When both modules are jointly enabled, their complementary effects result in the best overall performance, underscoring the importance of combining attention-level diagnostics with output-level guidance.

#### 4.4 Analysis and Discussion

**The Mean Variations of OIC and VAE Before and After DICA.** Figure 5 presents the average changes in VAE and OIC across shallow, middle, and deep layers before and after applying DICA. Overall, VAE consistently decreases, indicating more focused visual attention and reduced attention drift, while OIC increases across all layers, reflecting stronger reliance on image cues. These trends demonstrate that DICA effectively enhances visual grounding and improves the stability of the model’s reasoning process.

**Trigger Ratios and Triggered Layer Distribution of the Dual Module.** Figure 6 shows that VAE-GDC is triggered more frequently than OIC-

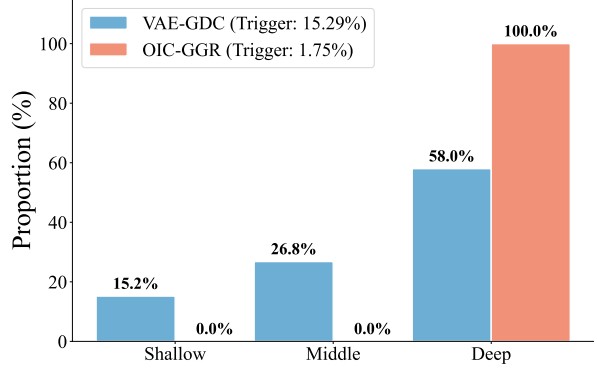


Figure 6: Trigger frequency and depth distribution of VAE-GDC and OIC-GGR on POPE.

Repair outcome	W→W	W→R	R→W	R→R
Overall (%)	8.2	4.9	1.3	86.9
Conditional (%)	62.6	37.4	1.5	98.5

Table 4: Repair success distribution on POPE. Here, W denotes wrong and R denotes right.

GGR, and both modules are primarily activated in deeper layers. This indicates that attention drift occurs more often than underuse of visual evidence, with both failure modes emerging mainly in later reasoning stages. The higher trigger frequency of VAE-GDC suggests that maintaining focused visual attention is more challenging for MLLMs than preserving general image dependence, especially as decoding progresses.

**Triggering Reliability and Correctness Transition.** Table 4, DICA achieves highly precise interventions, with an R→W rate of 1.3%, indicating hardly any degradation of correct predictions. At the same time, it corrects 37.4% of erroneous cases (W→R), confirming that VAE and OIC reliably identify hallucination risks and guide the model back to relevant visual evidence via contrastive alignment. DICA operates as an on-demand intervention mechanism that strikes a balance between robustness and hallucination mitigation.

## 5 Conclusion

This study identifies two prevalent failure modes in MLLM-based visual question answering: attention drift and underuse of visual evidence. We propose the DICA framework, which leverages two information-theoretic indicators, VAE and OIC, to dynamically monitor model behavior. During normal reasoning, VAE decreases and OIC remains high; abnormal VAE surges or OIC drops trigger hallucinations, which DICA corrects via VAE-

GDC and OIC-GGR, respectively. Extensive experiments across multiple models and benchmarks demonstrate that DICA improves accuracy, effectively mitigates hallucinations, and validates the effectiveness of the proposed dual-indicator guided alignment.

## Limitations

There are three main limitations to our work:

- Like other decoding-time hallucination mitigation methods, DICA introduces moderate computational overhead to enable real-time monitoring of visual grounding. This leads to a modest increase in latency compared to standard greedy decoding.
- The method relies on fixed threshold parameters, which may require manual adjustment for different models or downstream tasks, limiting its generalization across diverse MLLM architectures.
- DICA cannot compensate for perceptual blind spots inherent in the pretrained vision encoder or language model, and its performance is ultimately bounded by the base model’s visual and linguistic capacities.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051, and the Postgraduate Research and Innovation Foundation of Yunnan University under Grant No.KC-252513133. The authors would like to thank the anonymous reviewers for their constructive comments.

## References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, and reasoning. *arXiv preprint arXiv:2308.12966*.
- Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390.
- Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multi-modal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923.
- Xinlong Chen, Yuanxing Zhang, Qiang Liu, Junfei Wu, Fuzheng Zhang, and Tieniu Tan. 2025. Mixture of decoding: An attention-inspired adaptive decoding strategy to mitigate hallucinations in large vision-language models. *arXiv preprint arXiv:2505.17061*.
- 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*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.
- Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.
- Alessandro Favero, Luca Zancato, Matthew Trager, Sidharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.
- 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.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.
- Taehyeon Kim, Joonkee Kim, Gihun Lee, and Se-Young Yun. 2023. Instructive decoding: Instruction-tuned

- large language models are self-refiner from noisy instructions. *arXiv preprint arXiv:2311.00233*.
- Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- 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.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023d. [Evaluating object hallucination in large vision-language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.
- J. Lin. 1991. [Divergence measures based on the shannon entropy](#). *IEEE Transactions on Information Theory*, 37(1):145–151.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: Common objects in context](#). In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.
- Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. 2024a. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Kuanghong Liu, Jin Wang, Kangjian He, Dan Xu, and Xuejie Zhang. 2025a. Vision-aware multimodal prompt tuning for uploadable multi-source few-shot domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18897–18905.
- Qiang Liu, Xinlong Chen, Yue Ding, Bowen Song, Weiqiang Wang, Shu Wu, and Liang Wang. 2025b. Attention-guided self-reflection for zero-shot hallucination detection in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21016–21032.
- David Navon. 1977. [Forest before trees: The precedence of global features in visual perception](#). *Cognitive Psychology*, 9(3):353–383.
- Clement Neo, Shay B Cohen, and Fazl Barez. 2024. Interpreting context look-ups in transformers: Investigating attention-mlp interactions, 2024. *URL <https://arxiv.org/abs/2402.15055>*, page 25.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. [Object hallucination in image captioning](#). In *Proc. EMNLP*, pages 4035–4045. Association for Computational Linguistics.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

- Andrés Villa, Juan León Alcázar, Motasem Alfarrá, Vladimir Araujo, Alvaro Soto, and Bernard Ghanem. 2025. Eagle: Enhanced visual grounding minimizes hallucinations in instructional multimodal models. *arXiv preprint arXiv:2501.02699*.
- Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.
- Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2025. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1927–1951.
- Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. 2025. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1571–1590.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *Preprint, arXiv:2505.09388*.
- Hao Yang, Jin Wang, and Xuejie Zhang. 2025b. Evlmcot: Enhanced vision-language multi-cot for harmful meme detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 403–415. Springer.
- Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.
- Alfred L Yarbus. 2013. *Eye movements and vision*. Springer.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwan He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. Rlhv-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*, pages 13807–13816.
- Li Yuan, Yi Cai, Xudong Shen, Qing Li, Qingbao Huang, Zikun Deng, and Tao Wang. 2025. Collaborative multi-lora experts with achievement-based multi-tasks loss for unified multimodal information extraction. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6940–6948.
- Li Yuan, Qingfei Huang, Bingshan Zhu, Yi Cai, Qingbao Huang, Changmeng Zheng, Zikun Deng, and Tao Wang. 2026. Hybrid-dmkg: A hybrid reasoning framework over dynamic multimodal knowledge graphs for multimodal multihop qa with knowledge editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28032–28040.
- Yunhang Shen Yulei Qin Mengdan Zhang, Xu Lin Jinrui Yang Xiawu Zheng, Ke Li Xing Sun Yunsheng Wu, Rongrong Ji Chaoyou Fu, and Peixian Chen. 2021. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 18:2.
- Guangmin Zheng, Jun Kong, Jin Wang, and Xuejie Zhang. 2025. Enhanced multimodal chain-of-thought with visual self-contrastive distillation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.
- Guangmin Zheng, Jin Wang, Xiaobing Zhou, and Xuejie Zhang. 2024. Enhancing semantics in multimodal chain of thought via soft negative sampling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6059–6076.
- 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. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## A Proof of Metric Applicability

To validate that our proposed metrics, Visual Attention Entropy and Output-Image Correlation, faithfully reflect the human-like cognitive process of *Global Exploration*  $\rightarrow$  *Focal Exploitation*, we provide the following theoretical justifications based on information theory.

### 1. Entropy as an Indicator of Visual Exploration Scope

**Claim.** *The Visual Attention Entropy  $H^{(l,t)}$  serves as an information-theoretic proxy for the spatial scope of visual processing, characterizing the transition from global browsing to focal attention.*

*Proof.* Let  $\mathcal{V} = \{1, \dots, M\}$  be the set of visual token indices, and the attention distribution be  $\mathbf{p} \in \Delta^{M-1}$ . Consider two extreme cognitive states:

- **Global Browsing (Exploration):**  $p_i = 1/M$  for all  $i$ , entropy is maximized:

$$\text{VAE}(p) = - \sum_{i=1}^M \frac{1}{M} \log \frac{1}{M} = \log M. \quad (20)$$

- **Focal Attention (Exploitation):**  $p_k = 1$  for some token  $k$ , others 0, entropy is minimized:

$$\text{VAE}(p) = 0. \quad (21)$$

Since Shannon entropy (Shannon, 1948) is concave and uniquely maximized by the uniform distribution, a decreasing sequence  $\{H^{(l,t)}\}_t$  indicates a reduction in the effective receptive field, i.e., cognitive narrowing from scene-level to object-level focus.

### 2. OIC as a Measure of Visual Evidence Contribution

**Claim.** *The OIC quantifies the information gain from visual tokens over the textual prior, measuring reliance on visual evidence.*

*Proof.* Recall

$$\text{OIC}^{(l,t)} = \sum_v p_v \log \frac{p_v}{p_{\text{wi},v}}, \quad (22)$$

which is the Kullback-Leibler divergence (Kullback and Leibler, 1951)  $D_{\text{KL}}(P(\cdot | \text{Text, Image}) \| P(\cdot | \text{Text}))$ . From a Bayesian view, text is the prior, image is the observation:

$$\text{OIC}^{(l,t)} = \mathbb{E}_{v \sim P} \left[ \log \frac{P(v | \text{Text, Image})}{P(v | \text{Text})} \right]. \quad (23)$$

- If visual info is redundant or ignored:  $\text{OIC} \rightarrow 0$ .
- If visual info provides critical evidence:  $\text{OIC} \gg 0$ .

Hence, OIC directly measures the informational contribution of visual perception to the decision-making process.

### 3. Modeling the Cognitive Evolution

Based on the above analysis, the human VQA process can be described as a trajectory in (VAE, OIC) space:

$$\text{Process : } \begin{cases} \text{Phase I : } & t \text{ small, VAE}^{(t)} \approx \log M, \\ & \text{OIC}^{(t)} \text{ moderate,} \\ \text{Phase II : } & t \uparrow, \text{VAE}^{(t)} \downarrow, \\ & \text{context accumulates,} \\ \text{Phase III : } & t = t_{\text{ans}}, \text{VAE}^{(t)} \rightarrow 0, \\ & \text{OIC}^{(t)} \text{ maximized.} \end{cases}$$

This trajectory describes an idealized and interpretable pattern of cognitive evolution rather than a deterministic cognitive law.

## B Proof of Metric Validity

To understand the relationship between VAE and OIC, and how they reflect human-like cognitive processes in VQA, we analyze their theoretical connection through the lens of Information Theory and Transformer dynamics.

**OIC as Information Gain.** First, we observe that the definition of  $\text{OIC}^{(l,t)}$  is equivalent to the KL divergence between the image-conditioned output distribution  $\mathbf{p}^{(l,t)}$  and the text-only prior  $\mathbf{p}_{\text{wi}}^{(l,t)}$ :

$$\text{OIC}^{(l,t)} = D_{\text{KL}}(\mathbf{p}^{(l,t)} \| \mathbf{p}_{\text{wi}}^{(l,t)}). \quad (24)$$

This quantity measures the *information gain* provided by the visual context at step  $t$ . A high OIC implies that the visual features significantly alter the model's prediction away from the pure language prior.

**Attention Entropy as Feature Selectivity.** The visual context vector  $\mathbf{c}^{(l,t)}$  entering the feed-forward network is a weighted sum of visual tokens  $V = \{\mathbf{v}_i\}$ :

$$\mathbf{c}^{(l,t)} = \sum_{i:m_i=1} p_i^{(l,t)} \mathbf{v}_i. \quad (25)$$

We analyze two limiting cases of the attention entropy  $H^{(l,t)}$  to demonstrate the evolution process:

**Phase 1: High Entropy (Global Exploration).** During the initial stages of reasoning high VAE $^{(l,t)}$ , the attention distribution is dispersed, approximating a uniform distribution where  $p_i^{(l,t)} \approx 1/N_v$  where  $N_v$  is the number of visual tokens. The context vector becomes an average of the visual scene:

$$\lim_{\text{VAE} \rightarrow \text{VAE}_{\max}} \mathbf{c}^{(l,t)} \approx \frac{1}{N_v} \sum_i \mathbf{v}_i = \bar{\mathbf{v}}_{\text{global}}, \quad (26)$$

where  $\bar{\mathbf{v}}_{\text{global}}$  represents generic scene semantics. While this informs the global context, it rarely conflicts strongly with the language before specific tokens, resulting in a *moderate* divergence OIC. This mirrors the human cognitive phase of scanning for global semantics.

**Phase 2: Low Entropy (Focal Exploitation).** As the model resolves the semantics low VAE $^{(l,t)}$ , the attention distribution becomes peaked and sparse. In the limit where VAE $^{(l,t)} \rightarrow 0$ , the probability mass concentrates on a single critical token  $k$ :

$$\lim_{\text{VAE} \rightarrow 0} \mathbf{c}^{(l,t)} = \mathbf{v}_k. \quad (27)$$

In VQA, the visual token  $\mathbf{v}_k$  typically contains specific evidence (e.g., color, count, text) that is orthogonal to the text context. The injection of this highly specific feature  $\mathbf{v}_k$  causes the logits  $\mathbf{z}^{(l,t)}$  to shift significantly from  $\mathbf{z}_{\text{wi}}^{(l,t)}$ , maximizing the KL divergence:

$$\text{OIC}^{(l,t)} \propto \|\mathbf{z}^{(l,t)}(\mathbf{v}_k) - \mathbf{z}_{\text{wi}}^{(l,t)}\|. \quad (28)$$

This corresponds to the human cognitive phase of *focusing on key details to formulate an answer*.

**Conclusion.** Therefore, the temporal evolution from high entropy to low entropy, accompanied by a rise in OIC, mathematically formalizes the cognitive strategy:

$$\underbrace{\text{VAE}_{\text{high}}^{(l,t)} \xrightarrow{\text{time}} \text{VAE}_{\text{low}}^{(l,t)}}_{\text{Attention Focusing}} \implies \underbrace{\text{OIC}_{\text{low}}^{(l,t)} \xrightarrow{\text{time}} \text{OIC}_{\text{high}}^{(l,t)}}_{\text{Evidence Integration}}. \quad (29)$$

This inverse correlation validates that the model accurately mimics the human VQA process, starting with global perception and narrowing down to the extraction of specific visual evidence.

## C Related Work

**Multimodal Large Language Model.** Inspired by the success of LLMs (Touvron et al., 2023), recent research has extended their capabilities to the multimodal domain, leading to the rapid development of MLLMs (Liu et al., 2025b; Bai et al., 2023). With the introduction of LLMs, MLLMs have increasingly adopted unified generative frameworks (Liu et al., 2024b; Zhu et al., 2023), which significantly enhance reasoning, generalization, and adaptability across diverse vision-language tasks. Typically, MLLMs are trained via a two-stage paradigm, including vision-language feature alignment pretraining followed by instruction-based fine-tuning (Dai et al., 2023). Recent studies have further incorporated human preference optimization techniques, such as RLHF (Sun et al., 2024) and preference fine-tuning (Zhou et al., 2023). Despite these advances, MLLMs still suffer from pervasive hallucination issues, where generated responses deviate from visual evidence, posing critical challenges to their reliability in real-world applications. This limitation underscores the need to develop more robust mechanisms to enhance visual grounding and factual consistency in MLLMs.

**Hallucination in MLLMs.** Hallucinations, which generate content inconsistent with visual inputs or factual reality, have been widely studied in both language-only and vision-language settings (Yuan et al., 2026, 2025; Ji et al., 2023; Yao et al., 2023; Liu et al., 2024a; Li et al., 2023c; Yang et al., 2025b; Zheng et al., 2025, 2024; Liu et al., 2025a). Due to the requirement of precise visual-textual alignment, hallucinations in MLLMs are often more severe, particularly in tasks such as visual question answering and image captioning, where models tend to over-rely on linguistic priors when visual grounding is unreliable or ambiguous (Biten et al., 2022; Wang et al., 2024; Zhou et al., 2023).

Existing hallucination mitigation methods in MLLMs can be broadly categorized into two classes according to whether they require additional training. **Training-based approaches** aim to suppress hallucinations by modifying model architectures or training objectives. Typical strategies include supervised fine-tuning on carefully curated anti-hallucination datasets, reinforcement learning with hallucination-aware rewards, or contrastive and counterfactual data construction that explicitly penalizes hallucinated outputs (Liu et al., 2023a; Cai et al., 2024; Yu et al., 2024; Villa et al., 2025).

While effective in controlled settings, these methods are often costly in terms of data annotation and computation, and may compromise generalization when deployed beyond the training distribution.

In contrast, **training-free inference-time methods** attempt to regulate hallucinations without updating model parameters. A representative line of work leverages contrastive decoding, where hallucination-prone predictions are deliberately induced and contrasted against the original decoding stream to suppress spurious generations (Li et al., 2023b; Chuang et al., 2023). In the multimodal domain, such hallucination-inducing signals are constructed through noisy or removed visual inputs (Leng et al., 2024; Favero et al., 2024), adversarial or negative prompts (Woo et al., 2025; Kim et al., 2023), or attention manipulation and decoding heuristics (Huang et al., 2024). These approaches are attractive due to their flexibility and low deployment cost, and have shown promising empirical gains across benchmarks.

However, despite their differences, both training-based and training-free methods generally treat hallucination as a monolithic phenomenon and lack explicit mechanisms to diagnose its underlying causes. In particular, they fail to distinguish whether hallucinations arise from attention drift, or from underuse of visual evidence. As a result, existing methods tend to apply uniform correction strategies, such as globally suppressing language bias or enforcing stronger visual grounding, which may be misaligned with the actual failure mode and can even degrade correct predictions.

Motivated by this limitation, our proposed DICA departs from one-size-fits-all correction by explicitly diagnosing hallucination patterns during inference and performing targeted, adaptive interventions. By jointly monitoring VAE and OIC, DICA identifies whether hallucinations are caused by attention drift or underuse of visual evidence, and accordingly activates distinct mechanisms to recalibrate the reasoning process.

## D More Experimental Details

### D.1 Metrics

**Metrics on MME.** In the MME benchmark, each image  $i \in I$  is paired with two similar questions designed to elicit “yes” and “no” responses, denoted as  $\{q_i^y, q_i^n\}$ . Two metrics are used to evaluate model performance: the question-level accuracy (Acc) and the image-level accuracy (Acc<sup>+</sup>).

Table 5: CHAIR hallucination results on LLaVA-1.5.  $C_S / C_I$  (lower is better).

Method	LLaVA-1.5 (512)		LLaVA-1.5 (64)	
	$C_S$	$C_I$	$C_S$	$C_I$
Greedy	45.0	14.7	20.6	6.2
Nucleus	48.8	14.2	26.2	8.5
Beam Search	48.8	13.9	18.8	5.9
DoLA	47.8	13.8	20.4	6.3
OPERA	44.6	12.8	14.2	5.2
DICA	<b>41.8</b>	<b>11.5</b>	<b>10.4</b>	<b>4.3</b>

The question-level accuracy (Acc) is defined as:

$$\text{Acc} = \frac{\sum_{i \in I} \mathbb{I}[\text{MLLMs}(i, q_i^y) = \text{"Yes"}]}{|I| \times 2} + \frac{\sum_{i \in I} \mathbb{I}[\text{MLLMs}(i, q_i^n) = \text{"No"}]}{|I| \times 2}. \quad (30)$$

The image-level accuracy (Acc<sup>+</sup>) is defined as:

$$\text{Acc}^+ = \frac{\sum_{i \in I} \mathbb{I}[\text{MLLMs}(i, q_i^y) = \text{"Yes"}]}{|I|} + \frac{\mathbb{I}[\text{MLLMs}(i, q_i^n) = \text{"No"}]}{1}. \quad (31)$$

The final MME Score is computed as:

$$\text{MME Score} = \text{Acc} + \text{Acc}^+. \quad (32)$$

**Metrics on CHAIR.** The CHAIR benchmark evaluates hallucinations in image captioning using two metrics: CHAIR<sub>i</sub> and CHAIR<sub>s</sub>.

The instance-level hallucination rate is:

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \quad (12)$$

The sentence-level hallucination rate is:

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinations}\}|}{|\{\text{all sentences}\}|}. \quad (13)$$

Recall measures caption completeness:

$$\text{Recall} = \frac{|\{\text{accurately mentioned objects}\}|}{|\{\text{ground-truth objects}\}|}. \quad (14)$$

## D.2 Results

### D.2.1 Additional Results on CHAIR

Table 5 presents hallucination evaluation results on LLaVA-1.5 using the CHAIR benchmark, comparing various decoding strategies under two visual token settings: 512 (standard) and 64 (compressed). Lower CHAIR<sub>s</sub> and CHAIR<sub>i</sub> scores indicate fewer hallucinations. DICA consistently

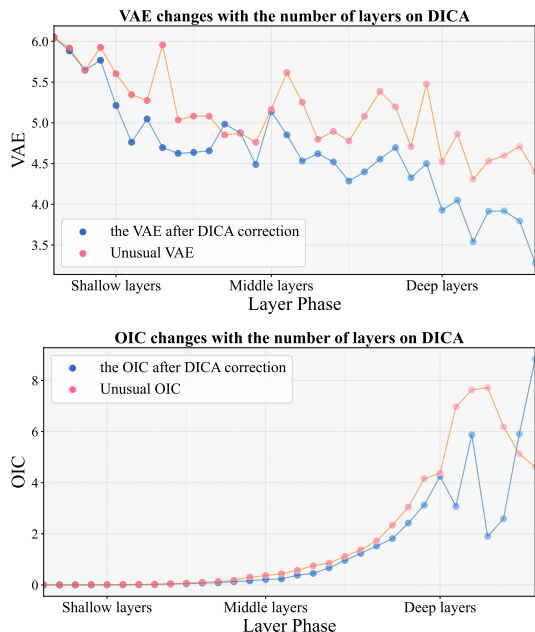


Figure 7: **Top:** VAE changes across layers phase on DICA. **Down:** OIC change across layers phase on DICA.

achieves the lowest hallucination rates among all methods, demonstrating its effectiveness in grounding model outputs in visual evidence. Notably, DICA maintains strong performance in both the 512-token and 64-token settings, indicating its robustness regardless of whether standard or compressed visual representations are used.

### D.2.2 Additional Results on POPE

Table 6 reports the complete experimental results on the POPE benchmark. DICA consistently outperforms nearly all existing methods, including VCD, M3ID, and other training-free and training-based baselines. On LLaVA-1.5 under the MS-COCO Popular setting, DICA achieves 87.7% accuracy, outperforming Sample (82.0%), VCD (82.1%), M3ID (82.8%), AVisC (78.2%), and MoD (85.7%). On Qwen-VL, DICA exceeds all competitors across all nine accuracy metrics. Overall, DICA achieves the strongest performance, particularly under adversarial and popular conditions, demonstrating superior visual grounding and effective hallucination mitigation during inference.

## D.3 More Analysis and Discussion

### D.3.1 VAE and OIC Performance Changes with DICA

Figure 7 illustrates the evolution of VAE and OIC across shallow, middle, and deep layers after applying the DICA intervention. The corrected tra-

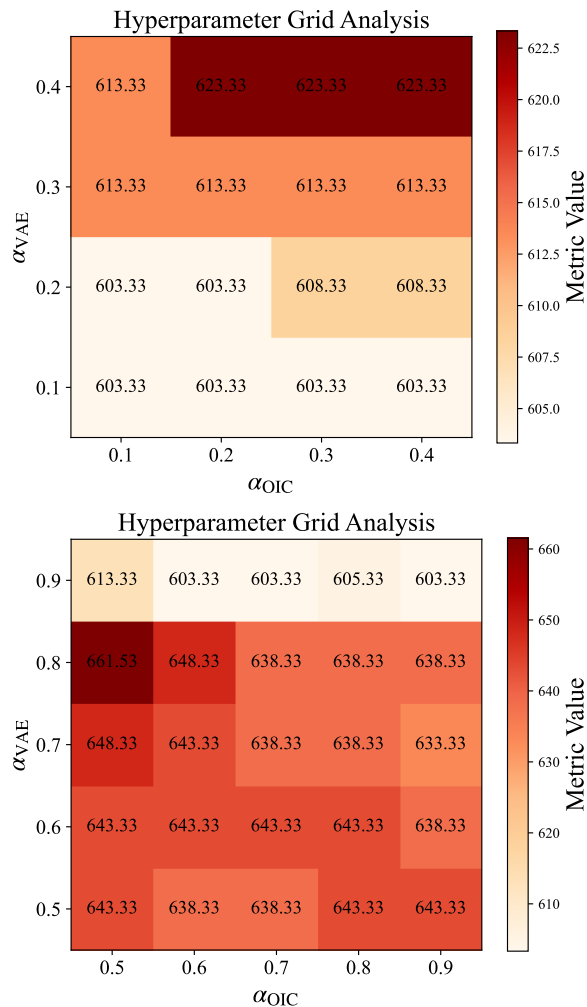


Figure 8: The upper figure corresponds to thresholds 0.1–0.4, and the lower figure to thresholds 0.5–0.9.

jectories demonstrate that VAE steadily decreases. Meanwhile, OIC consistently increases, reflecting a reasoning process that aligns with the human-like cognitive pattern of shifting from global exploration to focal exploitation. In contrast, the uncorrected anomalous steps reveal instances of attention drift, abnormally high VAE, and under-use of visual evidence, as well as abnormally low OIC. DICA effectively suppresses these anomalies, ensuring that visual attention remains focused and that model predictions stay grounded in image evidence throughout inference.

### D.3.2 Analysis of $\alpha_{VAE}$ and $\alpha_{OIC}$

The sensitivity of DICA to the drift and grounding thresholds  $\alpha_{VAE}$  and  $\alpha_{OIC}$  is evaluated through a grid search on the MME benchmark. Figure 8 top presents the results when both thresholds range from 0.1 to 0.4. In this low-threshold regime, performance remains at a relatively low level, with

Dataset	Setting	Method	LLaVA-v1.5				Qwen-VL			
			Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
MS-COCO	random	sampling	83.8	82.4	86.1	84.2	84.9	96.0	72.9	82.9
		VCD	85.0	82.7	86.1	84.2	85.5	96.0	71.1	83.6
		M3ID	86.1	83.2	86.8	85.0	85.3	95.1	74.2	83.4
		AvisC	82.3	78.4	89.3	83.5	82.9	96.2	68.5	80.0
		MoD	89.2	90.0	88.2	89.1	86.0	97.0	74.3	84.1
		DICA (Ours)	<b>89.2</b>	94.1	83.7	88.8	<b>87.4</b>	96.9	77.4	<b>86.0</b>
	popular	sampling	82.0	79.7	85.9	82.6	84.0	94.7	72.1	81.9
		VCD	82.1	78.5	88.3	83.2	84.9	94.5	74.9	83.6
		M3ID	82.8	80.1	88.5	84.1	84.2	94.1	73.8	82.7
		AvisC	78.2	72.7	90.3	80.5	82.8	95.5	68.9	80.1
		MoD	85.7	84.1	88.1	86.1	85.6	96.3	74.0	83.7
		DICA (Ours)	<b>87.7</b>	90.6	83.1	<b>86.7</b>	<b>87.5</b>	96.7	77.6	<b>86.1</b>
	adversarial	sampling	75.8	71.3	86.3	78.1	82.1	90.0	72.3	80.2
		VCD	76.3	71.5	87.3	78.7	84.0	90.6	74.9	82.0
		M3ID	77.1	71.8	87.6	78.9	83.2	90.4	73.1	80.8
		AvisC	74.2	68.4	89.9	77.7	81.2	91.9	68.5	78.5
		MoD	79.7	75.4	88.2	81.3	84.0	92.4	74.2	82.3
		DICA (Ours)	<b>83.2</b>	88.6	76.2	<b>82.0</b>	<b>84.8</b>	93.1	75.21	<b>83.7</b>
A-OKVQA	random	sampling	81.8	76.4	92.1	83.5	86.8	93.2	79.5	85.8
		VCD	81.2	75.2	93.0	83.2	87.4	92.9	81.1	86.6
		M3ID	82.9	76.8	94.1	84.6	87.1	92.4	80.3	85.9
		AvisC	79.1	71.9	95.5	82.1	84.7	93.0	74.9	83.0
		MoD	86.5	81.3	94.7	87.5	87.8	94.4	80.4	86.9
		DICA (Ours)	<b>89.2</b>	87.9	91.0	<b>89.4</b>	<b>89.5</b>	94.3	84.3	<b>88.9</b>
	popular	sampling	75.3	69.1	91.5	78.7	85.6	90.6	79.5	84.7
		VCD	74.7	68.2	92.5	78.5	86.3	89.5	81.2	85.1
		M3ID	75.8	69.8	92.1	79.4	85.9	90.7	79.2	84.6
		AvisC	71.8	64.7	95.6	77.2	83.9	90.9	75.5	82.4
		MoD	79.5	72.6	94.5	82.2	86.5	91.6	80.3	85.6
		DICA (Ours)	<b>88.6</b>	90.8	81.7	<b>86.0</b>	<b>88.2</b>	93.1	82.5	<b>87.5</b>
	adversarial	sampling	67.4	61.8	91.2	73.7	80.4	80.1	80.9	80.5
		VCD	68.1	61.9	93.8	74.6	80.7	80.1	81.6	80.8
		M3ID	68.3	62.1	93.4	74.6	80.5	80.7	80.2	80.4
		AvisC	64.4	58.8	96.1	73.0	78.1	80.1	74.7	77.3
		MoD	69.1	62.6	94.7	75.4	81.0	81.2	80.6	80.9
		DICA (Ours)	<b>77.8</b>	75.7	81.8	<b>78.6</b>	<b>81.3</b>	81.7	80.8	<b>81.3</b>
GQA	random	sampling	81.6	75.6	93.2	83.5	81.3	88.8	71.5	79.2
		VCD	82.2	76.0	94.1	84.1	82.0	87.6	74.5	80.5
		M3ID	83.3	76.8	94.0	84.5	82.4	88.1	72.8	79.7
		AvisC	79.0	71.4	96.7	82.2	80.5	89.9	68.6	77.8
		MoD	86.2	80.8	95.1	87.4	83.8	90.3	75.7	82.3
		DICA (Ours)	<b>89.1</b>	92.6	85.1	<b>88.7</b>	<b>86.7</b>	94.5	78.1	<b>85.4</b>
	popular	sampling	73.1	66.7	92.5	77.5	75.9	78.1	72.0	74.9
		VCD	71.5	64.7	94.5	76.8	75.9	76.6	74.7	75.6
		M3ID	72.3	64.9	94.8	77.1	76.8	78.9	75.2	77.0
		AvisC	67.4	60.9	97.1	74.8	74.2	77.9	67.5	72.3
		MoD	74.0	66.8	95.3	78.6	79.8	82.4	75.7	78.9
		DICA (Ours)	<b>81.9</b>	79.4	85.5	<b>82.5</b>	<b>82.1</b>	86.4	77.6	<b>81.7</b>
	adversarial	sampling	68.0	62.0	93.4	74.5	75.5	77.8	71.2	74.4
		VCD	67.6	61.5	94.4	74.5	76.7	77.8	74.7	76.2
		M3ID	67.2	61.0	93.9	74.0	77.1	78.6	74.9	76.7
		AvisC	64.1	58.5	96.7	72.9	75.5	80.5	67.4	73.4
		MoD	68.7	62.2	95.3	75.3	78.9	81.0	75.4	78.1
		DICA (Ours)	<b>78.4</b>	74.8	85.9	<b>79.9</b>	<b>79.2</b>	82.8	74.5	<b>78.2</b>

Table 6: Detailed evaluation results on POPE benchmark (Acc, Pre, Rec, F1) for LLaVA-v1.5 and Qwen-VL.

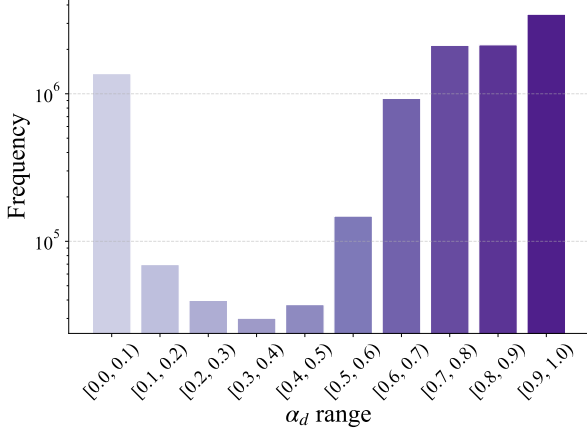


Figure 9: The change of  $\alpha_d$  On POPE.

metric values concentrated between 603.33 and 608.33, showing no clear improvement.

When the thresholds are set below 0.5, DICA becomes overly sensitive, causing the intervention mechanisms to be frequently triggered. Normal fluctuations and minor noise in the reasoning process are often misidentified as abnormal signals, leading VAE-GDC and OIC-GGR to apply unnecessary corrections. Such over-correction disrupts otherwise stable inference trajectories and degrades overall performance.

In contrast, the moderate-to-high threshold regime shown in the down part of Figure 8 effectively alleviates this issue. With  $\alpha_{\text{VAE}} = 0.8$  and  $\alpha_{\text{OIC}} = 0.5$ , interventions are activated only when substantial attention drift or insufficient visual grounding occurs. This selective activation avoids noise-induced corrections while enabling targeted adjustments, resulting in the best performance of 661.53. These results indicate that slightly higher threshold values strike a better balance between intervention frequency and correction strength, which is crucial for DICA to achieve optimal effectiveness.

### D.3.3 Analysis of $\alpha_d$

The value of  $\alpha_d$  exhibits a pronounced bimodal distribution during inference on the POPE benchmark Figure 9, with most values concentrated near 0 or 1. This shows that VAE-GDC’s contrastive correction is selectively activated, either fully engaged or nearly inactive—rather than applied uniformly. Values in  $[0.0, 0.1)$  occur when the Normal and Descriptive paths produce highly divergent outputs, typically during stable, visually grounded reasoning. Here, the Descriptive path which generates open-ended descriptions naturally differs from the

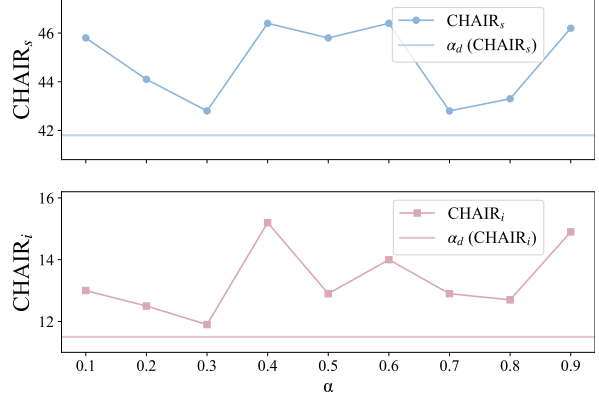


Figure 10: Performance with fixed  $\alpha$  on CHAIR.

task-specific Normal path, and the low  $\alpha_d$  prevents unnecessary correction, preserving fluency. Conversely, values in  $[0.9, 1.0)$  arise during attention drift: both paths yield diffuse, high-entropy distributions due to lost visual focus, making their outputs unusually similar. The near unity  $\alpha_d$  triggers strong contrastive decoding to recover a grounded prediction. The scarcity of intermediate values confirms that DICA acts as a sparse, decisive intervention, activating only when clear anomalies are detected.

We further fix  $\alpha_d$  to a constant value  $\alpha$  and conduct experiments on the CHAIR dataset. As shown in Figure 10, our results reveal a clear relationship between the blending weight  $\alpha$  and hallucination behavior. When  $\alpha$  is close to 0, the model heavily relies on the Normal path, which is tightly grounded in visual input for the specific task. This leads to strong visual alignment and the lowest CHAIR scores, indicating minimal object hallucination. As  $\alpha$  increases toward 1, the influence of the Descriptive path grows, introducing more open-ended and fluent language patterns but also increasing the risk of describing non-existent objects, as reflected by rising CHAIRs and CHAIRi. The best performance is consistently achieved at small  $\alpha$  values, confirming that precise visual grounding enabled by prioritizing the Normal path is crucial for hallucination suppression in captioning. This empirical trend validates the design of the adaptive  $\alpha_d$ , which automatically favors the Normal path during stable reasoning and only invokes the Descriptive path when attention drift necessitates corrective intervention.

### D.3.4 Analysis of $\beta_d$

The value of  $\beta_d$  exhibits a pronounced bimodal distribution during inference on the POPE bench-

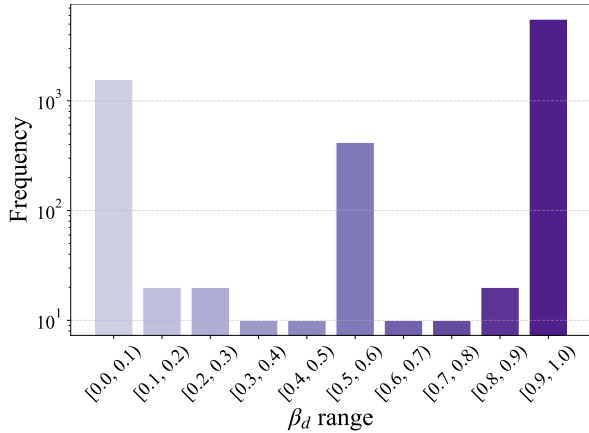


Figure 11: The change of  $\beta_d$  on POPE.

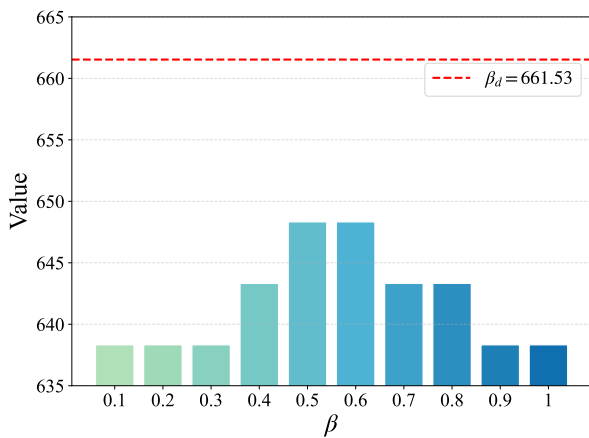


Figure 12: Performance with fixed  $\beta$  on MME.

mark. As shown in Figure 11, the majority of  $\beta_d$  values are clustered near the extremes of 0 and 1, rather than being uniformly distributed. Specifically, the high-density peak near 0 corresponds to stable reasoning phases where the model maintains a high correlation between its output and the relevant visual evidence, indicating that the internal attention mechanism is functioning correctly without excessive intervention. Conversely, the sharp peak near 1 signifies instances where the OIC drops significantly, reflecting potential attention drift or underuse of visual evidence. The fact that  $\beta_d$  effectively polarizes these two states demonstrates that DICA can act as a precise binary-like trigger. This adaptive mechanism ensures that contrastive alignment is only aggressively applied when a deviation from visual evidence is detected, thereby suppressing hallucinations while preserving the linguistic fluency of the model in normal states.

We fixed the original  $\beta_d$  value from the paper and replaced it with a unified constant hyperparameter  $\beta$ . As shown in Figure 12, the evalua-

tion on the MME benchmark shows that the original setting with  $\beta_d$  achieves a score of 661.53, which significantly outperforms all configurations using fixed  $\beta$  values. Specifically, across the tested range of  $\beta$  from 0.1 to 1.0, the best performance is only 648.33, which is clearly lower than the score obtained with  $\beta_d$ . This result demonstrates that the dynamic design of  $\beta_d$  more effectively balances generation fluency and visual grounding, leading to superior overall performance and further validating the effectiveness of the OIC-GGR module in DICA.

### D.3.5 Effectiveness of DICA in Visible

Without DICA, the model exhibits clear signs of **attention drift** and **underuse of visual evidence** as illustrated in Figure 3. For instance, when answering “Is there a sports ball in the image?”, the attention map becomes diffused in deeper layers, shifting away from the relevant object region; simultaneously, the log-probability of the correct answer drops sharply, leading the model to output an incorrect “No”. A similar failure occurs for the query “Is there a truck in the image?”, where visual cues are largely ignored despite their presence, resulting in hallucinated responses. “““ In contrast, after applying DICA see Figure 13, our dual-indicator monitoring detects these anomalies in the middle-to-late decoding stages: a sudden rise in VAE triggers VAE-Guided Drift Correction. Meanwhile, a drop in OIC activates OIC-Guided Grounding Restoration. As a result, the model’s attention is realigned to the semantically relevant regions, and the answer log-probabilities stabilize and converge toward the correct label. The final predictions become both visually grounded and accurate, demonstrating DICA’s effectiveness in restoring faithful cross-modal reasoning.

### D.3.6 Efficiency Analysis

We analyze the computational overhead introduced by DICA and compare it with the total time cost of alternative hallucination mitigation strategies, including training-based and data-engineering-based approaches. During inference, standard autoregressive decoding requires one forward pass per token. In contrast, DICA computes lightweight indicators from three computational streams: the normal path, the text-only path, and the descriptive path, which yields a theoretical upper bound of approximately three forward passes per token. However, the actual wall-clock latency increase is

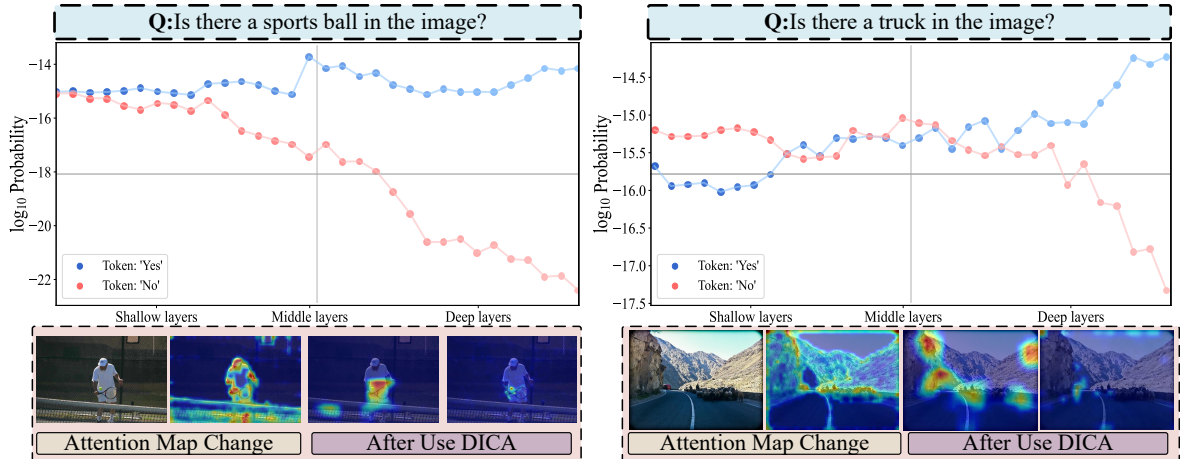


Figure 13: Evolution of answer log-probabilities and attention maps across layers applying DICA.

Approach	Training Time	Data Prep Overhead	Inference Cost	Deployment Agility
Retraining / RLHF	High	High	1×	Low
Data Engineering	Medium (fine-tuning required)	Very High	1×	Low
Beam Search ( $k = 5$ )	Zero	Zero	5×	High
<b>DICA (Ours)</b>	<b>Zero</b>	<b>Zero</b>	~2.5×	<b>Very High</b>

Table 7: Comparison of time and deployment costs across hallucination mitigation paradigms. Training Time is measured in GPU-hours. Data preparation overhead refers to the qualitative human and computational effort required to curate or generate additional supervision signals before training or inference. The inference cost is reported as a multiplicative factor relative to standard greedy decoding.

significantly smaller. This is because the vision encoder, the heaviest module in MLLMs, runs only once. Meanwhile, subsequent computations operate solely within the decoder layers and therefore benefit heavily from KV-cache reuse. Moreover, the additional streams are executed within the same decoding iteration. They are processed jointly on modern GPUs, whose parallel throughput leads to sub-linear latency growth relative to the number of streams. Another important observation is that DICA activates its computationally expensive correction modules only when the VAE or OIC indicators exceed their respective thresholds, meaning that the overhead is sparse rather than incurred at every decoding step. For example, on LLaVA-1.5, DICA adds roughly 0.04–0.08 seconds per token relative to greedy decoding. For typical VQA outputs under 20 tokens, the end-to-end delay remains below 1.5 seconds, which is well within the interactive latency tolerance.

The more substantial efficiency advantage of DICA appears when considering the full lifecycle cost of deploying a hallucination mitigation method. Unlike approaches requiring supervised fine-tuning, RLHF, or counterfactual data construction, DICA introduces zero training cost and zero

data preparation cost. In comparison, beam search incurs a 3–5× inference multiplier. Table 7 summarizes the comparison across methods. Despite a moderate increase in inference-time computation (approximately 2.5×), DICA eliminates the significantly higher cost associated with retraining or building large-scale, curated datasets.

### D.3.7 Performance on Qwen3

To further verify the generalization ability of our method, we conducted supplementary experiments on a more recent backbone model, Qwen3 (Yang et al., 2025a), under the POPE benchmark. Specifically, we compared the original Qwen3 baseline with Qwen3 enhanced by our DICA framework.

Table 8 presents the detailed results. Across all datasets and evaluation settings, DICA consistently improves both Accuracy and F1 score over the original Qwen3 model. Notably, performance gains are observed not only under Random splits but also under more challenging Popular and Adversarial settings, which are designed to evaluate robustness against hallucination biases.

These results indicate that DICA is not tightly coupled with a specific architecture but instead acts as a general and transferable hallucination mitiga-

Dataset	Setting	Qwen3(Base)		Qwen3(DICA)	
		Acc	F1	Acc	F1
MSCOCO	Random	89.50	88.67	91.73	91.15
	Popular	87.47	86.73	89.13	88.68
	Adversarial	85.70	85.11	87.23	86.95
AOKVQA	Random	92.80	92.71	94.10	94.07
	Popular	89.00	89.35	90.30	90.50
	Adversarial	82.70	84.14	84.37	85.52
GQA	Random	91.70	91.55	92.77	92.59
	Popular	87.60	88.01	89.37	89.48
	Adversarial	83.87	84.77	85.00	85.80

Table 8: Performance comparison on the POPE benchmark. Base denotes the original Qwen3 model.

tion framework.

## E More Case Studies

Figure 14 presents a qualitative comparison of different decoding and alignment methods on a representative indoor scene involving children and adults sharing food. As shown, baseline methods such as VCD, M3ID, VCD, and OPERA tend to introduce inconsistent or hallucinated details when generating descriptions, for example, misidentifying the event as a birthday party, adding non-existent decorations, or over-specifying social activities that are not visually grounded. These errors indicate that the models are overly influenced by spurious correlations or high-level priors rather than the actual visual evidence. In contrast, DICA produces a more faithful and conservative description that remains closely aligned with the image content, avoiding unsupported assumptions while preserving key visual elements. This case study demonstrates that DICA effectively mitigates attention drift and reduces semantic hallucination by correcting intermediate representations, leading to more reliable image-grounded generation.

Figure 15 presents a qualitative comparison of image description results generated by different decoding methods on the same input image. As shown, baseline methods tend to introduce varying degrees of hallucination or over-interpretation. Some outputs incorrectly infer specific cuisines, misidentify ingredients, or add contextual assumptions such as restaurant settings or instructional purposes that are not supported by the visual evidence. These errors indicate that the models are overly influenced by language priors, leading to semantic drift during decoding. In contrast, DICA produces a more faithful and conservative description, accurately reflecting the observable content in the image without introducing unsupported de-

tails. This case study highlights the effectiveness of DICA in suppressing hallucinated information and improving visual grounding.

Textual input: Please describe this image in detail.



Sample

The image shows a group of people gathered indoors around a long eating area, sharing a meal together. Both adults and children are present, suggesting a community or school-related event. **The environment appears similar to a cafeteria or classroom, where people are seated closely and interacting in a casual manner. The scene conveys a general sense of social dining and group interaction.**

VCD

This image depicts a lively birthday party taking place indoors, with many children gathered together to celebrate. **Colorful balloons and party decorations surround the group, enhancing the festive atmosphere.** The children are enjoying cake and snacks while adults take photos and supervise the celebration. The overall scene strongly resembles a cheerful birthday event.

M3ID

The image shows a group of children attending a special indoor celebration. **Various party elements such as desserts, gift items, and festive accessories can be observed.** The children appear excited as they sit together eating sweets, while adults document the moment. This scene likely represents a formal celebration organized for children.

AvisC

This image shows young children sitting on the floor in a classroom or daycare setting. **They appear to be participating in a structured group activity, possibly related to learning or storytelling.** Adults are nearby, likely acting as teachers guiding the session. The overall scene reflects an educational environment focused on group instruction.

MoD

The image captures a daycare environment where children are gathered indoors on the floor. **Some children appear to be eating snacks, while others seem to be playing or interacting with toys.** Adults supervise the children and manage different activities occurring at the same time. The scene combines elements of playtime and snack time.

OPERA

The image shows a group of children and adults indoors, interacting within a shared space. The individuals appear calm and engaged in a supervised group setting. **The environment looks organized and safe, suggesting a general group activity. No specific action or event is emphasized.**

DICA (ours)

The image shows a group of young children sitting closely together on colorful floor mats indoors, each eating slices of pizza from blue plates. **The setting resembles a daycare or children's activity room rather than a formal classroom. Several adults are standing or kneeling around the children, supervising them and taking photos.** The atmosphere is lively and informal, indicating a group meal or small children's gathering.

Figure 14: Case study of generative on CHAIR. We compare responses generated by sampling, VCD, M3ID, AvisC, MoD, and OPERA. Hallucinated content is highlighted red, while more detailed and accurate content is marked in green.

Textual input: Please describe this image in detail.



Sample

The image shows a plate of food placed on a wooden table. **The dish appears to contain rice along with various cooked ingredients mixed together.** Some vegetables and pieces of meat are visible, suggesting a balanced meal. **A glass of water can also be seen in the background, indicating a casual dining setting.**

VCD

This image depicts a Western-style dinner plate featuring **grilled steak served with mashed potatoes and mixed vegetables.** The meal appears to be professionally plated, resembling a restaurant-style dish. The warm tones and tableware further enhance the impression of a formal Western meal setting.

M3ID

The image shows a plated meal consisting of rice accompanied by curry sauce and vegetables. **The dish appears to be a form of curry rice, commonly found in South or Southeast Asian cuisine.** The sauce generously covers the rice, giving the meal a rich and flavorful appearance.

AvisC

This image shows a prepared meal used for demonstration or instructional purposes. **The food appears carefully arranged, suggesting it may be part of a cooking lesson or food presentation.** The plate includes rice and vegetables, arranged to showcase balance and portioning rather than casual dining.

MoD

The image presents a meal that could be either a homemade lunch or a restaurant dish. **It includes rice, vegetables, and possibly chicken or tofu.** The vegetables appear both steamed and fried, while the protein is unclear. The scene may represent either casual home dining or a plated food sample.

OPERA

The image shows a plate containing food placed on a table. **The contents are arranged neatly, and the environment appears calm and ordinary.** The scene represents a typical dining situation without emphasizing any specific type of cuisine or dish.

DICA (ours)

The image shows a plate of food placed on a wooden table, **consisting of a serving of white rice alongside a stir-fried dish made with sliced chicken and mixed vegetables such as broccoli and carrots.** **The food appears freshly cooked, with the rice kept separate from the stir-fry.** A glass of water is visible in the background, suggesting a casual dining setting, possibly at a restaurant or café.

Figure 15: Case study of generative on CHAIR. We compare responses generated by sampling, VCD, M3ID, AvisC, MoD, and OPERA. Hallucinated content is highlighted red, while more detailed and accurate content is marked in green.