# Dynamic Alignment Compensation for Hallucination Mitigation in Large Vision-Language Models

**Kairong Yu**[1,*] **Zixin Zhu**[1,*] **Le Yu**[2] **Hongwei Wang**[1,†]
[1]Zhejiang University, [2]Southeast University
[*]Equal contribution. [†]Corresponding author.
**Correspondence:** hongweiwang@intl.zju.edu.cn

## Abstract

Large Vision-Language Models (LVLMs) remain prone to hallucinations, producing responses that are irrelevant or inconsistent with the multimodal input. Existing mitigation methods mainly rely on external supervision, output calibration, or attention regulation, leaving the internal representation dynamics of autoregressive generation underexplored. We identify an inference-time failure mode in which cross-modal representations degrade across decoder layers and drift across generation steps, destabilizing token prediction and increasing hallucination risk. We propose *Dynamic Alignment Compensation* (DAC), a training-free inference-time method that detects representation divergence and selectively applies lightweight residual compensation. DAC combines Layer-wise Semantic Compensation to mitigate inter-layer degradation with Sequential Semantic Correction to constrain temporal drift. Experiments on nine hallucination-focused and general-purpose multimodal benchmarks across multiple LVLM backbones show that DAC consistently reduces hallucinations while maintaining strong overall performance.

## 1 Introduction

Large Vision-Language Models (LVLMs) have become a central paradigm for multimodal understanding, coupling visual inputs with autoregressive language generation (Koh et al., 2023; Tu et al., 2023). They have been widely applied to visual question answering (Antol et al., 2015; Shah et al., 2019; Jia et al., 2025), referring and reasoning-based segmentation (Lai et al., 2024; Lan et al., 2024; Zhu et al., 2025), and video retrieval (Wang et al., 2024a; Ventura et al., 2024). Despite their progress, LVLMs remain prone to hallucinations, producing responses that are irrelevant to or inconsistent with the given multimodal input (Gunjal et al., 2024; Lyu et al., 2025). Such failures undermine model reliability and can introduce substantial risks in high-stakes scenarios such as medical diagnosis (Lin et al., 2025) and autonomous driving (Ding et al., 2024).

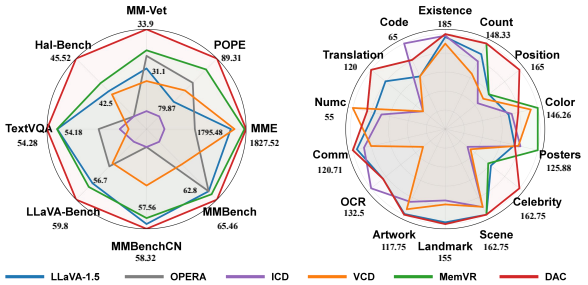



Figure 1: Performance comparison of different methods across multiple benchmarks. Hal-Bench denotes HallusionBench.

Existing hallucination mitigation methods can be broadly grouped into three paradigms. Retrieval-augmented methods incorporate external evidence during generation (Qu et al., 2025), while fine-tuning-based methods rely on additional supervision or domain-specific data (Yu et al., 2024). Both often require extra resources, such as retrieval databases, annotations, or model updates. Training-free methods instead intervene during inference without updating model parameters. Representative examples include logits-level contrastive decoding methods such as VCD and ICD (Leng et al., 2024; Wang et al., 2024c), attention-level methods such as OPERA (Huang et al., 2024), and representation-level methods such as MemVR (Zou et al., 2024).

In this work, we analyze LVLM hallucination through hidden-state dynamics during autoregressive inference. We identify two representation instabilities: inter-layer degradation across decoder layers and temporal drift across decoding steps. To mitigate them, we propose *Dynamic Alignment Compensation* (DAC), a training-free method that detects hidden-state divergence and applies lightweight residual compensation only



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 3392–3411
October 24-29, 2026 ©2026 Association for Computational Linguistics

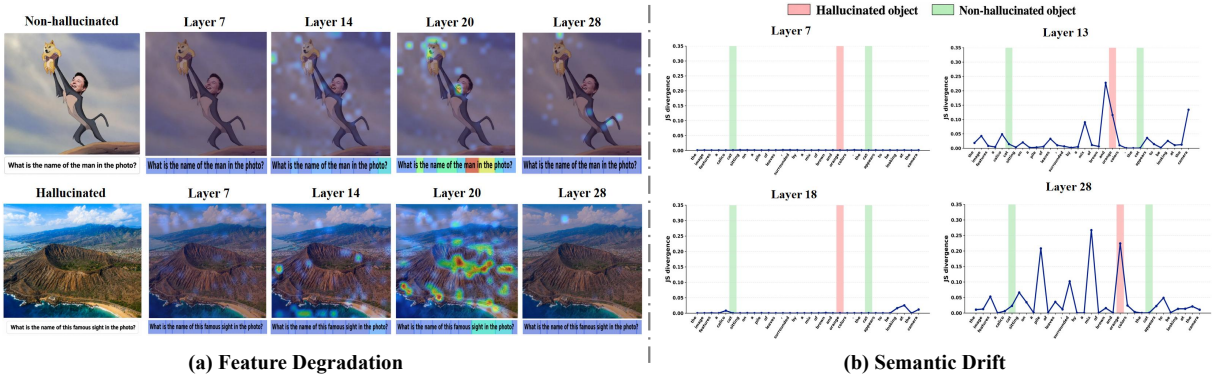



Figure 2: Layer-wise and temporal JS-divergence patterns in hallucinated and non-hallucinated LVLM generation. (a) Inter-layer JS divergence is projected onto image and text tokens at representative decoder layers. Hallucinated cases (bottom) show more frequent and stronger high-divergence regions than non-hallucinated cases (top), mainly in middle-to-late layers. (b) Across decoding steps, hallucinated object tokens exhibit larger temporal drift than non-hallucinated object tokens, with a similar middle-to-late-layer concentration.

when needed. DAC integrates Layer-wise Semantic Compensation (LSC) for inter-layer degradation and Sequential Semantic Correction (SSC) for temporal drift. Experiments on nine hallucination-focused and general-purpose multimodal benchmarks across multiple LVLM backbones show that DAC consistently reduces hallucinations while maintaining strong overall performance.

The main contributions of this work are:

- We identify inter-layer degradation and temporal drift as two inference-time representation instabilities in LVLM generation.
- We propose DAC, a training-free decoding method for dynamic hidden-state compensation along layer and temporal dimensions.
- We evaluate DAC on nine multimodal benchmarks across multiple LVLM backbones, showing consistent hallucination reduction without training or external resources.

## 2 Related Work

**Large Vision-Language Models.** Large Vision-Language Models (LVLMs) extend large language models (LLMs) to multimodal inputs by coupling visual encoders with autoregressive language models through cross-modal interfaces. Early systems such as BLIP-2 (Li et al., 2023a) and LLaVA (Liu et al., 2023) established this paradigm via query-based alignment and projection-based visual instruction tuning (Yin et al., 2023). Recent LVLMs further scale this design with stronger visual encoders, higher-resolution inputs, multi-image and video understanding, long-context modeling, and improved instruction alignment. Representative models include the LLaVA family (Liu et al., 2024a; Li et al., 2024), the Qwen-VL family (Wang et al., 2024b; Bai et al., 2025), InternVL (Chen et al., 2024), DeepSeek-VL2 (Wu et al., 2024), and Kimi-VL (Team et al., 2025). In this work, we use LLaVA-1.5 and Qwen-VL as the main backbones, and further evaluate DAC on Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen3-VL-4B, and LLaVA-NeXT-7B to examine its generality across model families, generations, and scales.

Despite this progress, LVLMs remain susceptible to hallucination, producing fluent and plausible responses that are inconsistent with the given multimodal context. This problem is particularly prominent in open-ended generation, where outputs are not constrained by a fixed answer space and local errors may accumulate during autoregressive decoding. Understanding the inference-time mechanisms behind such failures and developing reliable mitigation strategies therefore remain important challenges.

**Hallucination in LVLMs.** Hallucination in LVLMs has been studied from both diagnostic and mitigation perspectives (Yin et al., 2024; Zhou et al., 2023; Bai et al., 2024). Prior studies link hallucinations to imperfect fine-grained cross-modal grounding, object co-occurrence bias, and over-reliance on language priors, which may cause generated responses to become inconsistent with the given multimodal context (Rohrbach et al., 2018; Kim et al., 2023). Supervision-based methods mitigate hallucinations through hallucination-aware instruction tuning (Gunjal et al., 2024), post-hoc

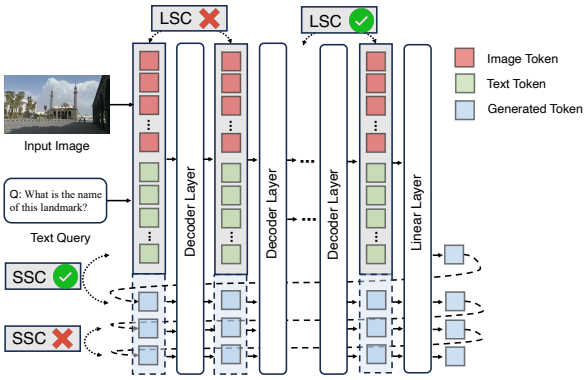



Figure 3: Overview of Dynamic Alignment Compensation (DAC). LSC compensates hidden states with large inter-layer representation divergence, while SSC constrains temporal drift of the current token using a layer-wise cache. Green checks denote triggered updates; red crosses denote unchanged states.

correction (Zhou et al., 2023), or reinforcement learning from human feedback (Yu et al., 2024). Although effective, they typically require additional annotations, auxiliary models, or costly training.

Training-free methods instead intervene during inference and can be categorized by their intervention targets. Attention-based methods, such as OPERA (Huang et al., 2024), EAH (Zhang et al., 2024), and AGLA (An et al., 2025), adjust attention distributions or visual-token usage during generation. Logits- and decoding-based methods, such as VCD (Leng et al., 2024), ICD (Wang et al., 2024c), SID (Huo et al., 2024), and DeGF (Zhang et al., 2025), calibrate token distributions through contrastive decoding, context-aware token selection, or generative feedback. Representation-based methods, such as MemVR (Zou et al., 2024), VISTA (Li et al., 2025), and AIR (Zhu et al., 2026), modify intermediate representations or salient internal features during generation. Complementary to these inference-time interventions, we study hallucination from the perspective of hidden-state dynamics, focusing on representation divergence along decoder depth and successive decoding steps.

## 3 Background and Motivation

### 3.1 Autoregressive Inference in LVLMs

We consider an LVLM $\mathcal{M}_\Theta$ parameterized by $\Theta$, consisting of a visual encoder, a cross-modal projection module, and a Transformer-based autoregressive decoder. Given an image $I$ and a textual query $Q$, the visual encoder and projection module convert $I$ into visual tokens $X_{\text{vis}}$, while $Q$ is tokenized into text tokens $X_{\text{txt}}$. The initial multimodal sequence is

$$X_0 = [X_{\text{vis}}, X_{\text{txt}}]. \tag{1}$$

At decoding step $t$, the decoder takes the extended sequence $X_t = [X_0, y_{<t}]$, where $y_{<t}$ denotes previously generated tokens. Let $H_{0,t}$ be the input embeddings of $X_t$. The hidden states are updated layer by layer as

$$H_{l,t} = \mathcal{F}_l(H_{l-1,t}), \quad l = 1, \ldots, L, \tag{2}$$

where $\mathcal{F}_l(\cdot)$ denotes the $l$-th decoder layer. The next-token distribution is computed from the final-layer hidden state at the current decoding position:

$$p_t = p_{\mathcal{M}_\Theta}(\cdot \mid I, Q, y_{<t}) = \text{softmax}(g_\Theta(h_{L,t})), \tag{3}$$

where $h_{L,t} = H_{L,t}^{(\text{cur})}$ and $g_\Theta(\cdot)$ is the output projection to the vocabulary space. The next token is then sampled or selected from $p_t$.

Under this formulation, multimodal context is carried forward through the evolution of intermediate hidden states. Therefore, representation instability during layer-wise propagation or across decoding steps can perturb the next-token distribution and increase the risk of hallucinated generation.

### 3.2 Representation Divergence in Hallucinated Generation

Prior studies have linked LVLM hallucination to dataset bias, object co-occurrence, and language priors (Zhou et al., 2023; Leng et al., 2024). We provide a complementary inference-time perspective by examining how hidden representations evolve during autoregressive decoding. Since next-token prediction depends on hidden states propagated through the decoder, abrupt representation changes along depth or decoding steps can perturb the decoding distribution. We use token-level Jensen-Shannon (JS) divergence between feature-normalized hidden states as a lightweight measure of such representation shifts.

**Inter-Layer Representation Degradation.** Decoder layers should refine multimodal representations while preserving semantic continuity. As shown in Fig. 2(a), non-hallucinated generation exhibits relatively sparse high-JS regions over image and text tokens, whereas hallucinated generation shows more frequent and stronger high-divergence regions. These regions are not uniformly distributed across depth; they mainly emerge in

middle-to-late decoder layers, while shallow and final layers show weaker separation. This pattern suggests that hallucinated generation is associated with localized inter-layer degradation during hidden-state propagation.

**Temporal Representation Drift.** Hallucination is also reflected in the temporal evolution of hidden states. Fig. 2(b) compares temporal JS divergence across decoding steps for hallucinated and non-hallucinated object tokens. Hallucinated object tokens exhibit larger representation drift, again concentrated in middle-to-late decoder layers. Together, these observations motivate DAC to compensate hidden-state instability along both layer and temporal dimensions.

## 4 Methodology

Motivated by the representation-divergence patterns in Sec. 3.2, we propose *Dynamic Alignment Compensation* (DAC), a training-free decoding method that stabilizes hidden-state propagation along both layer and temporal dimensions. DAC consists of two components: Layer-wise Semantic Compensation (LSC), which compensates token states with large inter-layer divergence, and Sequential Semantic Correction (SSC), which constrains the step-wise drift of the current-token representation.

### 4.1 Layer-wise Semantic Compensation

At decoding step $t$, let $H_{l,t} \in \mathbb{R}^{n_t \times d}$ denote the hidden states output by decoder layer $l$ for the current sequence $X_t = [X_0, y_{<t}]$, where $n_t$ is the sequence length and $d$ is the hidden dimension. For each token $i$ and layer $l \geq 2$, we normalize the adjacent-layer hidden states along the feature dimension:

$$p_{l-1,t}^{(i)} = s\left(H_{l-1,t}^{(i)}\right), \quad p_{l,t}^{(i)} = s\left(H_{l,t}^{(i)}\right), \tag{4}$$

where $s(\cdot)$ denotes feature-wise softmax.

For two normalized vectors $p$ and $q$, let $m = \frac{1}{2}(p+q)$. We define

$$\mathrm{JS}(p\|q) = \tfrac{1}{2}\mathrm{KL}(p\|m) + \tfrac{1}{2}\mathrm{KL}(q\|m). \tag{5}$$

The token-level inter-layer divergence is then computed as

$$D_{l,t}^{(i)} = \mathrm{JS}\left(p_{l-1,t}^{(i)} \,\|\, p_{l,t}^{(i)}\right). \tag{6}$$

LSC compensates only tokens whose divergence exceeds a threshold:

$$\bar{H}_{l,t}^{(i)} = H_{l,t}^{(i)} + \alpha\, \mathbf{1}\left[D_{l,t}^{(i)} > \gamma\right] H_{l-1,t}^{(i)}, \tag{7}$$

where $\gamma$ is the inter-layer divergence threshold, $\alpha$ controls the compensation strength, and $\mathbf{1}[\cdot]$ is the indicator function. Tokens below the threshold remain unchanged. The compensated state $\bar{H}_{l,t}$ is used as the layer output.

### 4.2 Sequential Semantic Correction

While LSC stabilizes hidden-state propagation across adjacent layers, autoregressive decoding may still introduce drift across successive generation steps. SSC constrains this drift using a lightweight layer-wise temporal cache. After LSC, let $\bar{h}_{l,t} = \bar{H}_{l,t}^{(\mathrm{cur})}$ denote the current-token hidden state at layer $l$. The cache $c_{l,t-1}$ stores the corrected current-token state from the previous decoding step. When the cache is available, we compute

$$q_{l,t} = s(\bar{h}_{l,t}), \quad q_{l,t-1} = s(c_{l,t-1}), \tag{8}$$

and measure temporal drift by

$$\Delta_{l,t} = \mathrm{JS}(q_{l,t} \,\|\, q_{l,t-1}). \tag{9}$$

If the drift exceeds a threshold, SSC moves the current state toward the cached state:

$$\tilde{h}_{l,t} = \bar{h}_{l,t} + \beta\, \mathbf{1}[\Delta_{l,t} > \tau]\left(c_{l,t-1} - \bar{h}_{l,t}\right), \tag{10}$$

where $\tau$ is the temporal drift threshold and $\beta \in [0,1]$ controls the correction strength. The corrected state $\tilde{h}_{l,t}$ is written back to the current-token position and stored as $c_{l,t}$ for the next decoding step. When no previous cache is available, SSC is skipped and the current state is directly cached.

At each decoding step, DAC applies LSC before SSC, so temporal correction operates on layer-stabilized hidden states. The complete inference pipeline is illustrated in Fig. 3 and summarized in Algorithm 1.

## 5 Experiments

### 5.1 Experiment Setup

We evaluate DAC on nine multimodal benchmarks covering both hallucination-focused and general-purpose evaluation. The hallucination-focused suite includes POPE (Li et al., 2023b), CHAIR (Rohrbach et al., 2018), and HallusionBench (Guan et al., 2024). The general-purpose suite includes MME (Fu et al., 2025),

Table 1: Performance evaluation on MME, MM-Vet, MMBench, TextVQA, and MMBench-CN benchmarks.

| Backbone | Method | MME | | | MM-Vet | MMBench | TextVQA | MMBench-CN |
|---|---|---|---|---|---|---|---|---|
| | | Perception↑ | Cognition↑ | Overall↑ | Score↑ | Accuracy↑ | Accuracy↑ | Accuracy↑ |
| LLaVA-1.5 | Regular | 1406.53 | 335.36 | 1741.89 | 31.10 | 62.80 | 54.18 | 57.56 |
| | OPERA | 1473.62 | 310.71 | 1784.34 | 32.00 | 62.80 | – | – |
| | ICD | 1341.63 | 295.36 | 1636.99 | 25.90 | 39.78 | 47.59 | 45.36 |
| | VCD | 1302.66 | 290.36 | 1593.02 | 30.20 | 54.21 | 45.71 | 51.46 |
| | MemVR | 1465.78 | 360.71 | 1826.49 | 32.40 | 63.75 | 54.22 | 56.62 |
| | VISTA | – | – | 1771.87 | – | – | – | – |
| | AGLA | 1472.66 | – | – | – | – | – | – |
| | DeGF | – | – | – | – | **65.50** | – | – |
| | AIR | – | – | – | 32.00 | – | – | – |
| | DAC | **1501.22** | **360.71** | **1861.93** | **33.90** | 65.46 | **54.38** | **58.32** |
| Qwen-VL-Chat | Regular | 1442.79 | 342.14 | 1784.93 | 49.00 | 56.53 | – | – |
| | OPERA | – | – | – | – | – | – | – |
| | ICD | 1472.06 | **361.69** | 1833.75 | 29.37 | 13.32 | – | – |
| | VCD | 1403.17 | 317.89 | 1721.06 | 34.54 | 39.18 | – | – |
| | MemVR | 1473.45 | 347.50 | 1820.95 | 49.60 | 56.44 | – | – |
| | DAC | **1509.16** | 354.64 | **1863.80** | **50.00** | **57.14** | – | – |

Table 2: Performance evaluation on the POPE benchmark.

| Method | Random | | Popular | | Adversarial | | Average | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | F1-score↑ | Accuracy↑ | F1-score↑ | Accuracy↑ | F1-score↑ | Accuracy↑ | F1-score↑ |
| LLaVA-1.5 | 83.49 | 82.28 | 79.98 | 79.34 | 76.03 | 76.26 | 79.83 | 79.29 |
| +OPERA | 87.53 | 86.45 | 84.21 | 83.50 | 80.88 | 80.69 | 84.21 | 83.55 |
| +ICD | 84.87 | 83.27 | 82.93 | 81.45 | 81.07 | 79.96 | 82.96 | 81.56 |
| +VCD | 86.84 | 86.83 | 82.65 | 83.37 | 77.31 | 79.28 | 82.27 | 83.16 |
| +MemVR | 88.50 | 87.34 | 87.10 | 86.01 | **85.20** | 84.28 | 86.93 | 85.88 |
| +AGLA | 88.54 | 87.71 | 85.14 | 84.68 | 81.13 | 81.36 | 84.94 | 84.58 |
| +DeGF | 89.03 | 88.74 | 86.63 | 86.28 | 81.63 | 80.59 | 85.76 | 85.65 |
| +VISTA | – | – | – | – | – | – | 86.15 | 86.29 |
| +DAC | **89.31** | **88.88** | **87.73** | **87.10** | 84.93 | **84.91** | **87.32** | **86.96** |

MM-Vet (Yu et al., 2023), MMBench (Liu et al., 2024b), MMBench-CN (Liu et al., 2024b), TextVQA (Singh et al., 2019), and LLaVA-Bench (Liu et al., 2023). Unless otherwise specified, all evaluations follow the official protocols. Detailed settings, metrics, and additional results are provided in Appendix A.

### 5.2 Results on General-Purpose Benchmarks

Tables 1 and 5 report results on general-purpose multimodal benchmarks. DAC consistently improves the vanilla decoding baseline on both LLaVA-1.5 and Qwen-VL-Chat across the reported evaluations. On MME, DAC increases the overall score by 120.04 points on LLaVA-1.5 and 78.87 points on Qwen-VL-Chat, with gains in both perception and cognition. It also improves MM-Vet and MMBench on both backbones, achieving 33.90/65.46 on LLaVA-1.5 and 50.00/57.14 on Qwen-VL-Chat, respectively. On LLaVA-1.5, DAC further improves TextVQA and MMBench-CN, indicating that hidden-state compensation does not compromise OCR-oriented or cross-lingual evaluation.

Table 5 further evaluates open-ended visual instruction following. DAC improves the average LLaVA-Bench score from 57.03 to 61.55 on LLaVA-1.5 and from 71.53 to 76.08 on Qwen-VL-Chat. Overall, DAC maintains strong general-purpose multimodal performance while remaining competitive with representative logits-, attention-, and representation-level mitigation methods.

### 5.3 Results on Hallucination Benchmarks

Tables 2, 4, and 7 report results on POPE, CHAIR, and HallusionBench. On POPE, DAC achieves the best average accuracy and F1, improving LLaVA-1.5 from 79.83 to 87.32 and from 79.29 to 86.96, respectively. It obtains the best results on the Random and Popular splits, and the highest F1 on the Adversarial split, showing consistent reduction of object-level hallucination.

On CHAIR, DAC reduces $C_s/C_i$ from 47.6/13.3 to 28.2/7.3 on LLaVA-1.5, and from 6.8/5.3 to 3.8/2.5 on Qwen-VL-Chat. Although recall decreases, the length-normalized analysis in Ap-

Table 3: Generalization to recent LVLM backbones on general-purpose multimodal benchmarks.

| Backbone | Method | MM-Vet (Score)↑ | MMBench (Acc.)↑ | TextVQA (Acc.)↑ | MMBench-CN (Acc.)↑ | LLaVA-Bench (Score)↑ |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | Regular | 62.1 | 78.35 | 77.41 | 77.06 | 82.37 |
| | +MemVR | 60.2 | 77.83 | 77.45 | 75.51 | 83.42 |
| | +ICD | 20.9 | 75.94 | 77.09 | 73.71 | 78.29 |
| | +VCD | 57.4 | 75.17 | 72.76 | 74.57 | 80.38 |
| | +DAC | **63.6** | **78.92** | **77.82** | **77.41** | **86.60** |
| Qwen2.5-VL-7B | Regular | 67.5 | 83.50 | 77.86 | 81.10 | 89.85 |
| | +MemVR | 68.6 | 83.33 | 75.65 | 80.49 | 89.22 |
| | +ICD | 19.4 | 81.87 | 75.65 | 79.37 | 81.73 |
| | +VCD | 65.7 | 81.52 | 76.02 | 79.55 | 85.04 |
| | +DAC | **69.3** | **83.91** | **77.94** | **81.37** | **95.18** |
| Qwen3-VL-4B | Regular | 68.0 | 84.10 | 78.13 | 81.27 | 95.62 |
| | +MemVR | 66.6 | 84.10 | 77.84 | 81.18 | 94.10 |
| | +ICD | 6.1 | 75.94 | 78.37 | 80.49 | 91.48 |
| | +VCD | 8.6 | 83.91 | 77.05 | 80.75 | 88.03 |
| | +DAC | **70.6** | **84.68** | **78.49** | **81.36** | **112.20** |
| LLaVA-NeXT-7B | Regular | 41.4 | 67.86 | 61.33 | 60.56 | 54.28 |
| | +MemVR | 40.3 | 67.96 | **61.41** | 60.56 | 57.07 |
| | +ICD | 19.2 | 62.80 | 60.28 | 55.49 | 49.82 |
| | +VCD | 25.8 | 63.40 | 54.64 | 56.26 | 52.56 |
| | +DAC | **41.9** | **68.44** | **61.41** | **60.85** | **58.88** |

Table 4: Performance evaluation on CHAIR. $C_s$ and $C_i$ denote sentence-level and instance-level hallucination rates, respectively; Len. denotes average response length.

| Method | $C_s \downarrow$ | $C_i \downarrow$ | Recall↑ | Len. |
|---|---|---|---|---|
| LLaVA-1.5 | 47.6 | 13.3 | 80.6 | 99.7 |
| +OPERA | 47.6 | 13.5 | 79.0 | 93.2 |
| +ICD | 56.2 | 16.3 | 16.3 | 103.4 |
| +VCD | 55.0 | 15.8 | 77.4 | 101.2 |
| +MemVR | 46.6 | 13.0 | **80.8** | 99.6 |
| +DAC | **28.2** | **7.3** | 72.0 | 88.9 |
| Qwen-VL-Chat | 6.8 | 5.3 | **53.4** | 17.6 |
| +OPERA | – | – | – | – |
| +ICD | 18.4 | 14.3 | 37.6 | 48.1 |
| +VCD | 13.0 | 12.3 | 47.9 | 115.7 |
| +MemVR | 4.8 | 3.3 | 52.3 | 15.0 |
| +DAC | **3.8** | **2.5** | 50.9 | 14.3 |

Table 5: Performance evaluation on LLaVA-Bench. C., D., Cp., and Avg. denote Conversation, Detail, Complex, and Average, respectively.

| Backbone | Method | C.↑ | D.↑ | Cp.↑ | All↑ | Avg.↑ |
|---|---|---|---|---|---|---|
| LLaVA-1.5 | Regular | 51.2 | 50.3 | 70.0 | 56.6 | 57.03 |
| | ICD | 49.3 | 46.9 | 63.1 | 54.9 | 53.55 |
| | VCD | 50.0 | 47.2 | 62.0 | 54.7 | 53.48 |
| | MemVR | 53.4 | 50.7 | 71.1 | 57.4 | 60.53 |
| | DAC | **57.5** | **55.8** | **73.1** | **59.8** | **61.55** |
| Qwen-VL-Chat | Regular | 68.9 | 61.6 | **82.3** | 73.3 | 71.53 |
| | ICD | 71.7 | 61.9 | 74.7 | 70.6 | 69.73 |
| | VCD | 33.8 | 23.9 | 30.2 | 29.6 | 29.38 |
| | MemVR | 69.4 | 62.8 | **82.3** | 73.7 | 72.05 |
| | DAC | **77.3** | **69.6** | 81.5 | **75.9** | **76.08** |

pendix B shows that DAC also lowers hallucinated object mentions per generated token, indicating that the improvement is not merely due to shorter outputs. On HallusionBench, DAC improves qACC from 8.13 to 17.05 and aACC from 41.5 to 45.5, while achieving the best easyA score. Overall, DAC reduces hallucinations across complementary evaluation protocols, suggesting that dynamic hidden-state compensation improves generation reliability beyond a single benchmark setting.

### 5.4 Generalization Across Recent LVLM Backbones and Scales

We further evaluate DAC on Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen3-VL-4B, and LLaVA-NeXT-7B to examine its transferability across recent LVLM families and model scales. As shown in Table 3, DAC improves the vanilla baseline on every reported general-purpose metric across the four backbones. On MM-Vet, DAC brings gains of 1.5, 1.8, 2.6, and 0.5 points, respectively. The improvements are more pronounced on LLaVA-Bench, where DAC increases the scores by 4.23, 5.33, 16.58, and 4.60 points. Compared with MemVR, ICD, and VCD, DAC achieves the best or tied-best result in every entry of Table 3, indicating strong generalization beyond the main backbones.

Table 6 further evaluates DAC on MME and POPE. DAC improves MME Overall by 9.85, 11.59, and 11.43 points on Qwen2.5-VL-3B, Qwen2.5-VL-7B, and LLaVA-NeXT-7B, respectively, while showing only a slight decrease on

Table 6: Comparison on MME and POPE across recent LVLM backbones.MME reports perception, cognition, and overall scores; POPE reports average accuracy and F1 over three splits. Best or tied-best results within each backbone block are bolded.

| Backbone | Method | MME(Perception)↑ | MME(Cognition)↑ | MME(Overall)↑ | POPE(Acc.)↑ | POPE(F1)↑ |
|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B | Regular | 1518.15 | 621.78 | 2139.93 | 87.38 | 86.20 |
| | +MemVR | 1499.00 | **625.35** | 2124.35 | 87.51 | 86.36 |
| | +ICD | 1473.30 | 571.42 | 2044.72 | 87.31 | 86.08 |
| | +VCD | 1512.03 | 595.35 | 2107.38 | 86.91 | 85.80 |
| | +DAC | **1528.00** | 621.78 | **2149.78** | **87.67** | **86.43** |
| Qwen2.5-VL-7B | Regular | 1692.78 | 624.28 | 2317.06 | 87.67 | 86.54 |
| | +MemVR | **1702.00** | 622.14 | 2324.14 | 87.68 | 86.56 |
| | +ICD | 1666.24 | 612.50 | 2278.74 | 87.73 | 86.68 |
| | +VCD | 1596.66 | 611.07 | 2207.73 | **88.05** | **87.13** |
| | +DAC | 1699.00 | **629.65** | **2328.65** | 87.80 | 86.73 |
| Qwen3-VL-4B | Regular | 1711.75 | **609.64** | **2321.39** | 89.53 | 88.95 |
| | +MemVR | **1720.99** | 581.42 | 2302.41 | 89.28 | 86.43 |
| | +ICD | 1699.48 | 596.42 | 2295.90 | 88.96 | 88.33 |
| | +VCD | 1682.90 | 606.78 | 2289.68 | 89.37 | 88.91 |
| | +DAC | 1711.75 | 608.21 | 2319.96 | **89.66** | **89.03** |
| LLaVA-NeXT-7B | Regular | 1519.29 | 322.50 | 1841.79 | 87.50 | 86.42 |
| | +MemVR | 1501.55 | 308.57 | 1810.12 | **87.71** | **86.70** |
| | +ICD | 1413.27 | 331.07 | 1744.34 | 86.25 | 85.09 |
| | +VCD | 1369.61 | **340.71** | 1710.32 | 86.06 | 85.06 |
| | +DAC | **1521.08** | 332.14 | **1853.22** | 87.45 | 86.37 |

Table 7: Performance evaluation on HallusionBench.

| Method | fACC↑ | qACC↑ | easyA↑ | hardA↑ | aACC↑ |
|---|---|---|---|---|---|
| LLaVA-1.5 | **17.9** | 8.13 | 36.0 | 36.7 | 41.5 |
| +OPERA | 16.2 | 5.49 | 37.6 | 35.4 | 41.2 |
| +ICD | 13.9 | 8.35 | 36.9 | 33.5 | 38.2 |
| +VCD | 13.9 | 11.4 | 33.0 | 34.7 | 41.1 |
| +MemVR | **17.9** | 9.01 | 36.9 | **37.7** | 42.5 |
| +DAC | 14.2 | **17.05** | **44.4** | 37.3 | **45.5** |

Table 8: Ablation results of LSC and SSC.

| LSC | SSC | Rec. | OCR | Kno. | Gen. | Spa. | Math | Total |
|---|---|---|---|---|---|---|---|---|
| × | × | 36.4 | 20.1 | 22.3 | 22.2 | 24.8 | 7.7 | 31.1 |
| × | ✓ | 37.2 | 23.5 | 23.3 | 22.6 | 28.9 | 7.7 | 33.0 |
| ✓ | × | 37.9 | 23.6 | 24.6 | **24.5** | 28.3 | 7.7 | 33.3 |
| ✓ | ✓ | **37.9** | **25.1** | **24.9** | 24.1 | **30.1** | **7.7** | **33.9** |

Qwen3-VL-4B. On POPE, DAC improves average accuracy/F1 by 0.29/0.23, 0.13/0.19, and 0.13/0.08 on the two Qwen2.5-VL variants and Qwen3-VL-4B, with nearly unchanged results on LLaVA-NeXT-7B. Overall, these results show that DAC remains effective across newer architectures and different parameter scales, supporting dynamic hidden-state compensation as a broadly applicable inference-time mechanism.

### 5.5 Ablation Studies

**Effect of LSC and SSC.** Table 8 ablates the two components of DAC on MM-Vet with LLaVA-1.5. Both components improve the vanilla baseline, raising the total score from 31.1 to 33.0 with SSC and to 33.3 with LSC. SSC mainly benefits OCR and spatial reasoning, while LSC yields broader gains across recognition, OCR, knowledge, generation, and spatial reasoning. Combining LSC and SSC achieves the best total score of 33.9, with the strongest OCR, knowledge, and spatial scores. These results suggest that layer-wise compensation and temporal correction provide complementary benefits for stabilizing hidden-state dynamics during inference.

**Hallucination-related JS divergence.** The case visualization in Fig. 2 provides an intuitive example of the representation instability associated with hallucinated generation. In the layer-wise view, hallucinated generation exhibits denser and stronger high-JS regions over both image and text tokens than non-hallucinated generation, with the most visible changes appearing in middle-to-late decoder layers. The temporal view shows a consistent trend: tokens corresponding to hallucinated objects undergo larger step-wise drift than tokens corresponding to non-hallucinated objects. These observations motivate a dataset-level analysis to examine whether such patterns persist beyond individual examples.

We conduct this analysis on CHAIR with LLaVA-1.5. For inter-layer divergence, Fig. 4(a) reports the high-JS ratio, which measures how frequently tokens exhibit large changes between adjacent decoder layers. The outliers are concentrated

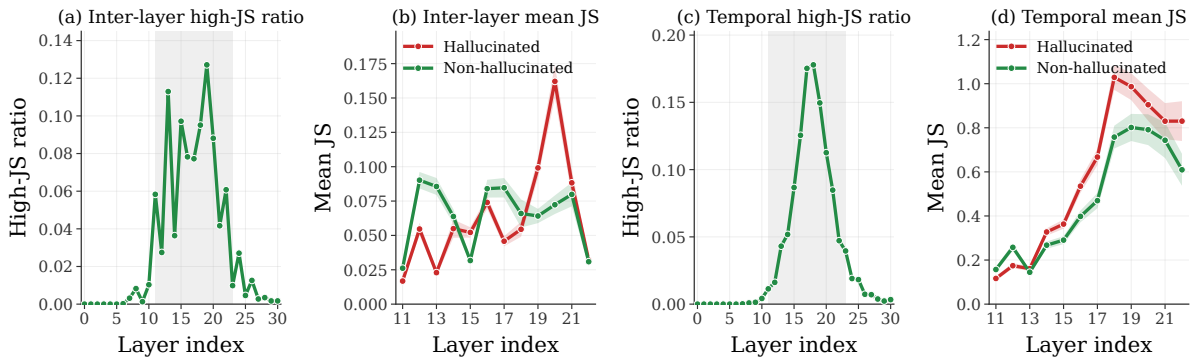



Figure 4: JS-divergence diagnostics on CHAIR with LLaVA-1.5. (a),(b) show the inter-layer high-JS ratio and mean JS across decoder layers, while (c),(d) show the corresponding temporal statistics across decoding steps. Mean-JS curves compare hallucinated and non-hallucinated object tokens. Gray regions highlight the middle-to-late layers where high-JS outliers concentrate, and shaded bands denote standard error.

Table 9: Inference efficiency on MM-Vet with LLaVA-1.5. Best or tied-best values are bolded.

| Method | Latency (ms/token)↓ | Throughput (tokens/ms)↑ | T/80 Tokens (ms)↓ | Memory (MB)↓ |
|---|---|---|---|---|
| Greedy | **36.8** | **0.027** | **2946.91** | 13797 |
| Sampling | 37.1 | **0.027** | 2972.99 | 13797 |
| MemVR | **36.8** | **0.027** | 2949.12 | 13800 |
| ICD | 79.0 | 0.013 | 6320.52 | **13778** |
| OPERA | 137.1 | 0.008 | 11501.57 | 21110 |
| VCD | 76.1 | 0.013 | 6085.18 | 13781 |
| DAC | 37.1 | 0.026 | 2975.45 | 13797 |

in middle-to-late layers rather than being uniformly distributed across depth. Fig. 4(b) further compares the mean JS of hallucinated and non-hallucinated object tokens within this region. Hallucinated object tokens show a clear late-layer peak, indicating not only more frequent instability but also stronger representation shifts.

The temporal statistics show the same tendency. Fig. 4(c) measures high-JS ratios between consecutive decoding steps and again identifies middle-to-late layers as the main unstable region. Fig. 4(d) shows that hallucinated object tokens have consistently larger temporal JS divergence than non-hallucinated ones, suggesting that hallucinated generation is associated with accumulated drift along the decoding trajectory. Together, the case-level visualization and dataset-level diagnostics support the design of DAC: LSC targets localized inter-layer instability, while SSC constrains temporal drift during autoregressive decoding.

**Inference efficiency.** We further evaluate the inference cost of DAC on MM-Vet with LLaVA-1.5. As shown in Table 9, DAC introduces negligible overhead over standard decoding. Compared with greedy decoding, DAC increases per-token latency only from 36.8 to 37.1 ms/token, while keeping the same peak memory usage of 13,797 MB. Its throughput remains close to standard decoding, decreasing slightly from 0.027 to 0.026 tokens/ms. For generating 80 tokens, DAC takes 2975.45 ms, which is comparable to greedy decoding and sampling, and substantially faster than ICD, VCD, and OPERA. Specifically, DAC reduces the 80-token generation time by 52.9%, 51.1%, and 74.1% compared with ICD, VCD, and OPERA, respectively. These results show that DAC improves generation reliability with little additional inference cost, without relying on extra decoding branches or auxiliary models.

## 6 Conclusion

We introduced *Dynamic Alignment Compensation* (DAC), a training-free inference-time method for mitigating LVLM hallucinations by stabilizing hidden-state dynamics. DAC detects inter-layer representation degradation and temporal representation drift, and applies lightweight compensation through Layer-wise Semantic Compensation and Sequential Semantic Correction. Experiments on nine hallucination-focused and general-purpose multimodal benchmarks show that DAC consistently reduces hallucinations while maintaining strong overall performance across multiple LVLM backbones and model scales. Further analysis confirms the complementarity of its two components and shows that DAC incurs negligible inference overhead. These results highlight dynamic hidden-state compensation as a practical direction for improving the reliability of LVLM generation.

## 7 Limitations

This work focuses on training-free hallucination mitigation for image-based LVLMs. Although DAC is evaluated on multiple hallucination-focused and general-purpose benchmarks across several open-source backbones, its effectiveness on closed-source models, video or multi-image inputs, long-context multimodal generation, and highly specialized domains remains to be further studied. DAC uses feature-wise softmax and Jensen-Shannon divergence as lightweight signals for detecting representation instability. While these signals are empirically effective, other distance measures or adaptive criteria may further improve robustness. In addition, we use fixed thresholds and compensation coefficients across all datasets and models to avoid dataset-specific tuning; future work may explore input-adaptive parameter selection. Finally, DAC reduces hallucination risk but does not guarantee factual correctness. For high-stakes applications, additional verification and safety assessment are still required.

## Acknowledgments

This work was supported by the National Key Research and Development Program of China (2024YFF0907803 and 2024YFF0907802), the National Natural Science Foundation of China (62276230), and the Research Fund for International Scientists of the National Natural Science Foundation of China (72350710798).

## References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. 2024. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13668–13677.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

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

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.

Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. 2025. Vqa2: visual question answering for video quality assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6751–6760.

Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. 2023. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Pro-*

*ceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2585–2595.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2023. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36:21487–21506.

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

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. 2024. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*.

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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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.

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

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

Qika Lin, Yifan Zhu, Xin Mei, Ling Huang, Jingying Ma, Kai He, Zhen Peng, Erik Cambria, and Mengling Feng. 2025. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *Information Fusion*, 116:102795.

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 *European conference on computer vision*, pages 740–755. Springer.

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. 2025. Realrag: Retrieval-augmented realistic image generation via self-reflective contrastive learning. *arXiv preprint arXiv:2502.00848*.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jiashuo Sun, Daizong Liu, and Jianfeng Dong. 2025. Alleviating hallucination in large vision-language models with active retrieval augmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(9):1–22.

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

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Haoqin Tu, Bingchen Zhao, Chen Wei, and Cihang Xie. 2023. Sight beyond text: Multi-modal training enhances llms in truthfulness and ethics. *arXiv preprint arXiv:2309.07120*.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5270–5279.

Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2024a. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16551–16560.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024c. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, and 1 others. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36:26650–26685.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. 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*, pages 13807–13816.

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

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

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

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*.

Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. 2025. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3686–3696.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, and Hanwang Zhang. 2026. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. *arXiv preprint arXiv:2602.24041*.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and 1 others. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

## A Benchmarks and Metrics

**Datasets and metrics.** We evaluate the proposed method on both hallucination-focused and general-purpose multimodal benchmarks. The hallucination-focused suite includes CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023b), and HallusionBench (Guan et al., 2024), which assess object hallucination, visual grounding, and image-context reasoning from complementary perspectives. To examine whether hallucination mitigation preserves general multimodal capability, we further evaluate on MME (Fu et al., 2025), MMBench and MMBench-CN (Liu et al., 2024b), MM-Vet (Yu et al., 2023), LLaVA-Bench (Liu et al., 2023, 2024a), and TextVQA (Singh et al., 2019). Unless otherwise specified, all results follow the official evaluation protocols of the corresponding benchmarks.

**CHAIR.** CHAIR evaluates object hallucination in image captioning by comparing generated object mentions with the ground-truth objects present in the image. We use the MSCOCO val2014 split (Lin et al., 2014), which provides annotations for 80 object categories, and randomly sample 500 images for evaluation. Each LVLM is prompted with "Please describe this image in detail." CHAIR reports hallucination at both the object-instance and sentence levels:

$$\begin{aligned} \text{CHAIR}_i &= |\mathcal{O}_h|/|\mathcal{O}_m|, \\ \text{CHAIR}_s &= |\mathcal{S}_h|/|\mathcal{S}|. \end{aligned} \tag{11}$$

Here, $\mathcal{O}_h$ denotes hallucinated object mentions, $\mathcal{O}_m$ denotes all mentioned objects, $\mathcal{S}_h$ denotes generated sentences containing hallucinated objects, and $\mathcal{S}$ denotes all generated sentences. Lower CHAIR scores indicate fewer hallucinations.

**POPE.** POPE is a binary VQA-style benchmark designed to evaluate object hallucination in LVLMs. Each question asks whether a specific object appears in the image, typically in the form "Is [object] in this image?" To enforce a clear binary response format, we append a yes-or-no instruction to each question. Following the standard protocol, queried objects are sampled under random, popular, and adversarial settings, and results are reported separately for each setting.

**HallusionBench.** HallusionBench evaluates hallucination-related failures in LVLMs under challenging image-context reasoning scenarios. It focuses on two major error sources: language hallucination, where the model relies on linguistic priors that conflict with visual evidence, and visual illusion, where misleading visual cues lead to incorrect predictions. The benchmark contains 346 images and 1,129 expert-written questions, enabling fine-grained evaluation of whether model responses are grounded in both the image content and the textual prompt.

**MME.** MME provides a comprehensive evaluation of multimodal perception and cognition. The perception split covers visual understanding tasks such as object existence, counting, spatial localization, color recognition, scene understanding, landmark recognition, artwork identification, and OCR. The cognition split evaluates higher-level abilities, including commonsense reasoning, numerical calculation, translation, and code-related reasoning. Each sample is formulated as a binary question, and the model is required to answer yes or no.

**MMBench and MMBench-CN.** MMBench evaluates general visual understanding and reasoning under a unified multiple-choice VQA protocol. It covers a broad range of abilities, including object and attribute recognition, spatial relation understanding, scene comprehension, commonsense reasoning, and text-aware visual reasoning. MMBench-CN follows the same evaluation framework in Chinese, providing a complementary assessment of cross-lingual visual grounding and instruction following. We report accuracy under the official evaluation setting.

**MM-Vet.** MM-Vet assesses open-ended multimodal reasoning across diverse capability dimensions, including spatial understanding, attribute and relation reasoning, compositional reasoning, multi-step inference, and text-rich visual reasoning. Since answers are free-form, model outputs are evaluated by an automatic judging pipeline that compares generated responses with reference answers and assigns graded scores according to correctness and instruction following.

**LLaVA-Bench.** LLaVA-Bench evaluates visual instruction-following ability in open-ended scenarios, including multi-turn dialogue, detailed image description, and visual reasoning. It contains 60 image-question pairs sampled from natural images. Model responses are compared with reference answers and scored according to the benchmark protocol, providing an additional measure of whether the proposed method preserves general instruction-following ability.

**TextVQA.** TextVQA evaluates the ability of LVLMs to read and reason over scene text in real-world images. Questions often require extracting text from signs, menus, product labels, posters, and other text-rich regions, and then grounding the extracted text in the visual context. We report the standard VQA-style accuracy based on answer matching.

## B Length-Normalized CHAIR Analysis

The standard CHAIR results in Table 4 show that DAC substantially reduces object hallucination. Since DAC also produces slightly shorter captions, we further examine whether this reduction is merely caused by decreased output length. To this end, we report length-normalized hallucinated-object rates on CHAIR.

For the $n$-th generated caption, let $h_n$ denote the number of hallucinated object mentions and $T_n$ denote the number of generated tokens. We define the corpus-level micro rate as

$$\mathrm{HObj@100}_{\mu} = 100 \cdot \frac{\sum_{n=1}^{N} h_n}{\sum_{n=1}^{N} T_n}. \tag{12}$$

The sample-level macro rate is defined as

$$\mathrm{HObj@100}_{M} = \frac{100}{N} \sum_{n=1}^{N} \frac{h_n}{T_n}. \tag{13}$$

Here, $\mathrm{HObj@100}_{\mu}$ measures hallucinated object mentions per 100 generated tokens over the whole evaluation set, while $\mathrm{HObj@100}_{M}$ first computes this rate for each sample and then averages across samples.

Table 10: Length-normalized CHAIR analysis on LLaVA-1.5. HObj@$100_\mu$ measures hallucinated object mentions per 100 generated tokens at the corpus level, while HObj@$100_M$ averages this rate over samples.

| **Method** | **HObj@$100_\mu$↓** | **HObj@$100_M$↓** |
|---|---|---|
| Regular | 1.1206 | 1.0314 |
| **DAC** | **0.5797** | **0.6569** |

As shown in Table 10, DAC reduces both normalized hallucination rates. HObj@$100_\mu$ decreases from 1.1206 to 0.5797, and HObj@$100_M$ decreases from 1.0314 to 0.6569. These results indicate that the CHAIR improvement is not simply an artifact of shorter outputs. Instead, DAC more effectively suppresses unsupported object mentions after controlling for generation length.

## C Backbones, Baselines, and Evaluation Protocol

**Backbones.** We evaluate DAC on both established and recent open-source LVLM backbones. The main experiments use LLaVA-1.5 (Liu et al., 2023) and Qwen-VL-Chat as representative backbones, following common practice in hallucination-mitigation evaluation. To further assess generality across newer architectures and model scales, we evaluate DAC on Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen3-VL-4B (Wang et al., 2024b; Bai et al., 2025), and LLaVA-NeXT-7B (Liu et al., 2024a). This set covers different model families, generations, and parameter scales, allowing us to examine whether DAC is tied to a specific backbone or remains effective for stronger recent LVLMs.

**Baselines.** We compare DAC with regular greedy decoding and representative training-free hallucination mitigation methods. VCD (Leng et al., 2024) and ICD (Wang et al., 2024c) are logits- or decoding-level contrastive methods that calibrate token distributions during generation. OPERA (Huang et al., 2024) is an attention-level method that mitigates hallucination through over-trust penalty and retrospection allocation. MemVR (Zou et al., 2024) is a hidden-state-based method that performs visual retracing in memory space. These baselines cover the major classes of training-free interventions, including logits-level, attention-level, and representation-level mitigation. For completeness, we also include directly comparable reported results from recent methods such as AGLA (An et al., 2025), DeGF (Zhang et al., 2025), VISTA (Li et al., 2025), and AIR (Zhu et al., 2026) when available. Entries that are unsupported by a method or incompatible with our implementation environment are marked as unavailable.

For all methods reimplemented in our experiments, we follow the official implementations and recommended hyperparameters whenever possible. Regular decoding and DAC use greedy decoding unless otherwise specified. Baseline-specific decoding configurations are provided in the reproducibility section. OPERA is omitted on Qwen-based backbones because its implementation depends on older versions of PyTorch and Transformers that are incompatible with our Qwen evaluation pipeline.

**Evaluation protocol.** We evaluate all models on nine multimodal benchmarks: POPE (Li et al., 2023b), CHAIR (Rohrbach et al., 2018), HallusionBench (Guan et al., 2024), MME (Fu et al., 2025), MM-Vet (Yu et al., 2023), MMBench and MMBench-CN (Liu et al., 2024b), TextVQA (Singh et al., 2019), and LLaVA-Bench (Liu et al., 2023, 2024a). These benchmarks cover hallucination-focused evaluation as well as general multimodal perception, reasoning, OCR, cross-lingual understanding, and open-ended instruction following. For each benchmark, we use the questions from the official annotation files and only adapt the surface format to the chat template required by each LVLM, such as inserting the corresponding system roles, user roles, and image tokens. No benchmark-specific prompt engineering is introduced for DAC.

For POPE, we follow the official COCO-based protocol and report results under the Random, Popular, and Adversarial splits. For CHAIR, we sample images from the MSCOCO Val2014 split (Lin et al., 2014) and query each model with the fixed prompt, "Please describe this image in detail," so that object hallucinations can be measured from free-form captions. For MMBench and MMBench-CN, we follow the official multiple-choice evaluation setting and report accuracy on the corresponding development splits. MME is evaluated under its standard yes-or-no protocol, with results reported for both perception and cognition. TextVQA is evaluated using the standard VQA-style answer-matching accuracy.

For open-ended benchmarks, we follow the official judging pipelines. MM-Vet scores are ob-

tained through the official online evaluator. LLaVA-Bench and HallusionBench are evaluated using their GPT-based judging protocols, and the same evaluator setting is applied to all methods evaluated by us within each benchmark. This ensures that performance differences primarily reflect decoding behavior rather than differences in evaluation prompts or judging procedures.

## D Reproducibility

**Implementation details.** Unless otherwise specified by a baseline method, we use greedy decoding for regular decoding and DAC. For all benchmarks, questions from the official annotation files are used as prompts, and only the surface format is adapted to the chat template required by each LVLM, such as inserting the corresponding system roles, user roles, and image tokens. No benchmark-specific prompt engineering is introduced for DAC. All benchmark evaluations follow the official protocols unless otherwise stated.

For POPE, we evaluate on the COCO-based split and report results under the Random, Popular, and Adversarial settings. For CHAIR, we use images sampled from the MSCOCO Val2014 split and query each model with the fixed prompt, "Please describe this image in detail." The same image subsets are used across all compared methods. We sample three image sets with different random seeds and report the averaged results. For MM-Vet, we use the official online evaluator. For LLaVA-Bench (in-the-wild) and HallusionBench, we follow the corresponding GPT-based judging pipelines and apply the same evaluator setting to all methods evaluated by us.

**Reimplemented baselines.** For all reimplemented baselines, we follow the official implementations and recommended hyperparameters whenever available. VCD (Leng et al., 2024) is implemented with visual contrastive decoding over the original and distorted visual inputs. Following the official implementation, we set `do_sample=True`, `temperature=1`, `cd_alpha=1`, `cd_beta=0.1`, and `noise_step=500`.

For ICD (Wang et al., 2024c), we follow the official contrastive decoding configuration. The contrastive penalty is set to $\lambda = 1$, and the adaptive plausibility constraint is set to $\alpha_{\text{ICD}} = 0.1$. We use sampling from the modified softmax distribution with `temperature=1`, `top_p=1`, repetition penalty 1, and beam size 1.

For OPERA (Huang et al., 2024), we use the official beam-search setting with `do_sample=False`, beam size 5, `scale_factor=50`, `threshold=15`, `num_attn_candidates=5`, and `penalty_weights=1`. OPERA is omitted on Qwen-based backbones because its implementation depends on older versions of PyTorch and Transformers that are incompatible with our Qwen evaluation pipeline.

For MemVR (Zou et al., 2024), we follow the official visual-retracing configuration with `retracing_ratio=0.12`, `entropy_threshold=0.75`, `starting_layer=5`, and `ending_layer=16`. We use greedy decoding with `do_sample=False`, `temperature=0`, and beam size 1 unless the official script specifies otherwise.

**Reported baselines.** In addition to the methods reimplemented above, we include directly comparable reported results from recent training-free mitigation methods, including AGLA (An et al., 2025), DeGF (Zhang et al., 2025), VISTA (Li et al., 2025), and AIR (Zhu et al., 2026), when results are available under the same backbone and benchmark protocol. These entries are taken from the corresponding papers or official result tables without additional re-tuning. Missing entries indicate that the method does not support the corresponding setting or that directly comparable results are not publicly available. These reported baselines are not included in the inference-efficiency comparison.

**DAC configuration.** For DAC, we use greedy decoding with `do_sample=False`, `temperature=0`, and beam size 1, following the regular decoding setting. The DAC hyperparameters are fixed for all datasets: $\gamma = 0.5$ and $\alpha = 0.2$ for Layer-wise Semantic Compensation, and $\tau = 0.3$ and $\beta = 0.3$ for Sequential Semantic Correction. We do not perform dataset-specific hyperparameter tuning.

**Computational resources.** All experiments are conducted on NVIDIA RTX 4090 GPUs. For efficiency evaluation, all methods are tested under the same hardware environment using LLaVA-1.5 on MM-Vet, and we report per-token latency, decoding throughput, total time for generating 80 tokens, and peak GPU memory usage.

## E The Algorithm of DAC

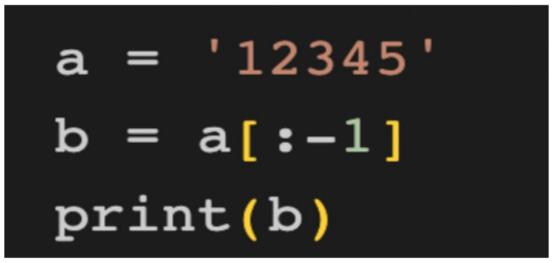
In this section, we present the full algorithmic details of DAC.

**Algorithm 1** Dynamic Alignment Compensation (DAC)

**Require:** LVLM $\mathcal{M}_\Theta$ with decoder blocks $\{\mathcal{F}_l\}_{l=1}^{L}$
**Require:** Image $I$, query $Q$
**Require:** Thresholds $\gamma, \tau$; coefficients $\alpha, \beta$
**Ensure:** Generated sequence $y_{1:T}$

```
Obtain visual tokens X_vis from I
Obtain textual tokens X_txt from Q
X_0 ← [X_vis, X_txt]
Initialize temporal caches c_l ← ∅, l = 1, ..., L
Let s(·) denote feature-wise softmax
for t = 1 to T do
    X_t ← [X_0, y_<t]
    H_{0,t} ← Embed(X_t)
    for l = 1 to L do
        H_{l,t} ← F_l(H_{l-1,t})
        if l > 1 then
            for each token i in X_t do
                p^(i)_{l-1,t} ← s(H^(i)_{l-1,t})
                p^(i)_{l,t} ← s(H^(i)_{l,t})
                D^(i)_{l,t} ← JS(p^(i)_{l-1,t} || p^(i)_{l,t})
                if D^(i)_{l,t} > γ then
                    H^(i)_{l,t} ← H^(i)_{l,t} + α H^(i)_{l-1,t}
                end if
            end for
        end if
        h_{l,t} ← H^(cur)_{l,t}
        if c_l ≠ ∅ then
            q_{l,t} ← s(h_{l,t})
            q_{l,t-1} ← s(c_l)
            Δ_{l,t} ← JS(q_{l,t} || q_{l,t-1})
            if Δ_{l,t} > τ then
                h_{l,t} ← h_{l,t} + β(c_l − h_{l,t})
                H^(cur)_{l,t} ← h_{l,t}
            end if
        end if
        c_l ← h_{l,t}
    end for
    y_t ← Decode(H^(cur)_{L,t})
end for
```

# F Examples of Capability Integrations




The image shows a python code. Is the output of the code '1234'? Please answer yes or no.

LLaVA Answer: No.
LLaVA(+DAC) Answer: Yes.
Ground Truth: Yes.

Figure 5: A case study comparing the levels of hallucination among various baselines

Here is a picture of the cake I cut. Did I cut it at least once? Please answer yes or no.

LLaVA Answer: Yes.
LLaVA(+DAC) Answer: No.
Ground Truth: No.

Figure 6: A case study comparing the levels of hallucination among various baselines

Are there four people in this image? Please answer yes or no.

LLaVA Answer: No.
LLaVA(+DAC) Answer: Yes.
Ground Truth: Yes.

Figure 7: A case study comparing the levels of hallucination among various baselines

Is this a photo of Liuyang Confucius Temple? Please answer yes or no.

LLaVA Answer: No.
LLaVA(+DAC) Answer: Yes.
Ground Truth: Yes.

Figure 8: A case study comparing the levels of hallucination among various baselines

Are there any red shoes in the image? Please answer yes or no.

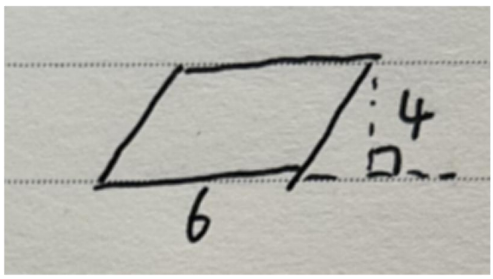
LLaVA Answer: No.
LLaVA(+DAC) Answer: Yes.
Ground Truth: Yes.

Figure 9: A case study comparing the levels of hallucination among various baselines

Is the blue umbrella above the black umbrella? Please answer yes or no.

LLaVA Answer: Yes.
LLaVA(+DAC) Answer: No.
Ground Truth: No.

Figure 10: A case study comparing the levels of hallucination among various baselines




Is the area of the parallelogram in the picture equal to 6? Please answer yes or no.

LLaVA Answer: Yes.
LLaVA(+DAC) Answer: No.
Ground Truth: No.

Figure 11: A case study comparing the levels of hallucination among various baselines

Is this picture captured in a place of rock arch? Please answer yes or no.

LLaVA Answer: No.
LLaVA(+DAC) Answer: Yes.
Ground Truth: Yes.

Figure 12: A case study comparing the levels of hallucination among various baselines

Is this artwork displayed in museo di san salvi, florence? Please answer yes or no.

LLaVA Answer: Yes.
LLaVA(+DAC) Answer: No.
Ground Truth: No.

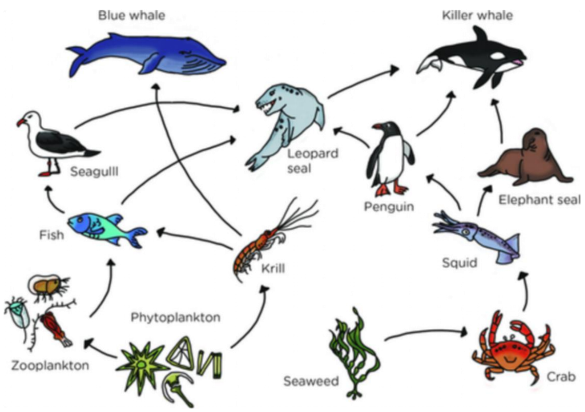
Figure 13: A case study comparing the levels of hallucination among various baselines

Is this movie originated from the country or region of uk? Please answer yes or no.

LLaVA Answer: No.
LLaVA(+DAC) Answer: Yes.
Ground Truth:Yes.

Figure 14: A case study comparing the levels of hallucination among various baselines

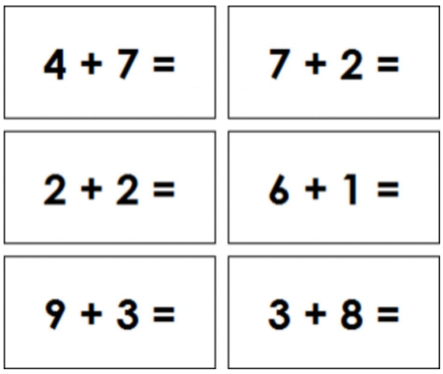
**Question:** What is the calories for the entire bag?
**Ground Truth:** 275.
**Required Capabilities:** OCR; Math




**Question:** Which are producers in this food web?
**Ground Truth:** Phytoplankton and Seaweed.
**Required Capabilities:** OCR; Knowledge; Spatial Awarenes

Figure 15: A case study comparing the levels of hallucination among various baselines

**Question:** Where is this photo taken?
**Ground Truth:** Qatar.
**Required Capabilities:** Recognition; Knowledge




**Question:** What is the answer to the bottom left equation
**Ground Truth:** 12.
**Required Capabilities:** OCR; Math; Spatial Awareness

Figure 16: A case study comparing the levels of hallucination among various baselines

**Question:** What is the lamp on, a side table or a nightstand?
**Ground Truth:** Side table.
**Required Capabilities:** Recognition; Spatial Awareness

**Question:** Is the person bigger than the car?
**Ground Truth:** No.
**Required Capabilities:** Recognition; Knowledge; Spatial Awareness

Figure 17: A case study comparing the levels of hallucination among various baselines