DiVE: Decoupling Intra-layer Visual Evidence for Mitigating Hallucinations in Large Vision-Language Models

Anonymous ACL submission

Abstract

Recent Large Vision-Language Models (LVLMs) have achieved significant progress yet frequently suffer from visual hallucinations, often stemming from an over-reliance on language priors rather than visual evidence. Existing decoding-based approaches often rely on input perturbations to weaken language priors, but they do not explicitly decouple visual evidence from mixed vision–language representations. To address these limitations, we propose **DiVE** (**D**ecoupling **i**ntra-layer **V**isual **E**vidence). DiVE dynamically identifies layers enriched with visual information and performs intra-layer decoupling to extract aggregated visual evidence. By suppressing this evidence to construct a language-prior-dominated reference distribution, DiVE employs contrastive decoding to calibrate the output logits, thereby mitigating hallucinations. Extensive experiments across diverse LVLM architectures demonstrate that DiVE achieves state-of-the-art performance among decoding-based methods on multiple benchmarks. Crucially, it eliminates the latency of an extra forward pass, offering a lightweight and efficient solution.

1 Introduction

Recent advances in LVLMs have led to substantial improvements in multimodal reasoning (Bai et al., 2025; Liu et al., 2023; Dai et al., 2023). Despite these advancements, current LVLMs frequently suffer from visual hallucinations, where they generate plausible but factually incorrect content that contradicts visual inputs (Leng et al., 2024; Li et al., 2023). This phenomenon primarily stems from an over-reliance on language priors and statistical biases acquired during pre-training (Rohrbach et al., 2018; Leng et al., 2024; Ma et al., 2023). Essentially, when visual evidence are subtle or overwhelmed, LVLMs effectively revert to acting as "blind" language models, allowing language priors to dominate generation. (Min et al., 2025; Wu et al., 2025).

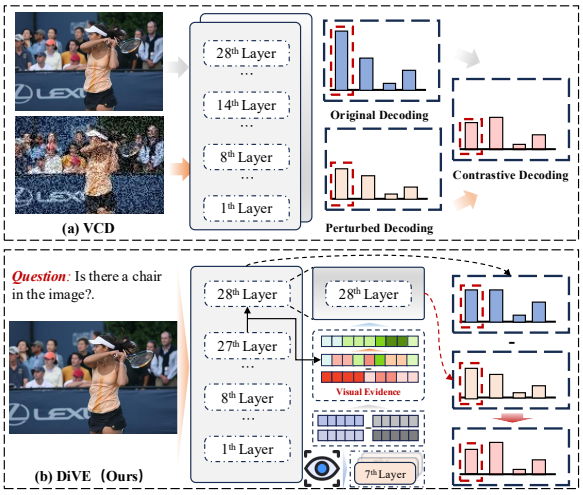


Figure 1: Comparison of perturbation-based contrastive decoding strategy (e.g., VCD) and our proposed DiVE.

To mitigate visual hallucinations, as shown in Figure 1 (a), inference-time decoding strategies (e.g., VCD (Leng et al., 2024), ICD (Wang et al., 2024) and M3ID (Favero et al., 2024)) contrast predictions against those from perturbed inputs. However, these methods inherently incur a 2× latency penalty by requiring dual forward passes. Furthermore, their reliance on perturbations creates an unstable contrastive reference, often introducing noise that degrades the precision of hallucination mitigation.

To reduce the computational overhead of perturbation-based methods (Leng et al., 2024; Wang et al., 2024), recent works focus on internal manipulation within a single forward pass. Methods such as OPERA (Huang et al., 2024) and AVISC (Woo et al., 2025) leverage attention patterns or weights to calibrate logits. Meanwhile, approaches like SID (Huo et al., 2025) and ONLY (Wan et al., 2025) construct contrastive references by utilizing differences between specific

intermediate layers. Nevertheless, these methods operate on mixed multimodal features without decoupling visual evidence from textual context. This entanglement limits their ability to suppress hallucinations when strong language priors overwhelm visual evidence.

To further *avoid perturbation-based references without extra passes* and *decouple visual evidence for calibration against language priors*, we propose DiVE (**D**ecoupling **i**ntra-layer **V**isual **E**vidence). As shown in Figure 1 (b), unlike VCD (Figure 1 (a)) which relies on an external noisy image to construct a contrastive reference, DiVE decouples visual evidence directly from the intra-layer attention. The framework proceeds in three steps. First, we perform adaptive layer identification to locate visual-evidence layers via structural constraints and entropy metrics. Second, we decouple the attention output to extract the effective visual contribution by subtracting the underlying textual attention bias. Finally, we apply contrastive decoding, leveraging the aggregated visual evidence to calibrate logits against language priors, thereby mitigating visual hallucinations.

We validate DiVE on multiple LVLM backbones and standard hallucination benchmarks. DiVE achieves consistent and significant hallucination mitigation, outperforming strong training-free decoding-time baselines. Notably, DiVE eliminates the latency of an extra forward pass and introduces only lightweight computation during standard decoding.

Our main contributions are summarized as follows:

- We propose **DiVE**, a training-free decoding framework that builds an internal contrastive reference by identifying visual-evidence layers, and decoupling intra-layer visual evidence for visual hallucination mitigation.
- Extensive experiments across multiple LVLM backbones and hallucination benchmarks show that DiVE achieves state-of-the-art hallucination mitigation, without requiring an extra forward pass.
- DiVE is highly adaptable and can be easily adapted to a wide range of LVLM backbones with different architectures, highlighting its potential for broad applicability.

2 Related Work

Hallucination in LVLMs. Despite strong multimodal reasoning capabilities (Liu et al., 2023; Bai et al., 2025; Dai et al., 2023), LVLMs frequently suffer from visual hallucinations—generating content contradictory to visual inputs (Li et al., 2023; Fu et al., 2025; Rohrbach et al., 2018). This phenomenon primarily stems from an over-reliance on language priors and spurious correlations established during pre-training (Rohrbach et al., 2018; Leng et al., 2024). While mitigation strategies via instruction tuning (Liu et al., 2024) or alignment (RLHF or DPO) (Sun et al., 2024; Yu et al., 2024; Xie et al., 2024) are effective, they are data-intensive and computationally expensive. Similarly, post-hoc correctors introduce significant latency (Zhou et al., 2023; Yin et al., 2024; Zhao et al., 2024) or rely on external models (Sahu et al., 2024; Wan et al., 2024). Consequently, training-free inference-time methods for hallucination mitigation have received growing attention.

Decoding-Time Hallucination Mitigation. Decoding-time strategies aim to mitigate hallucinations without parameter updates. A prominent paradigm is contrastive decoding (Leng et al., 2024; Wang et al., 2024; Favero et al., 2024), which penalizes language priors by contrasting logits against perturbed inputs. However, this inherently doubles computational overhead. To enhance efficiency, internal manipulation methods are adopted to exploit intrinsic model behaviors. Following layer-contrasting approaches like DoLa (Chuang et al., 2024), methods such as OPERA (Huang et al., 2024) and SID (Huo et al., 2025) leverage attention patterns or layer discrepancies. Notably, ONLY (Wan et al., 2025) suggests that single-layer intervention is sufficient, and AVISC (Woo et al., 2025) employs attention-based calibration. Nevertheless, existing methods mainly steer decoding using attention-derived cues (e.g., OPERA and AVISC) or other internal intervention signals(e.g., SID and ONLY); these methods lack a clear distinction between visual evidence and the textual context, which limits their robustness when language priors dominate.

3 Methodology

In this work, we present **DiVE**, a training-free decoding framework for hallucination mitigation. As illustrated in Figure 2, DiVE extracts intra-layer visual evidence, suppresses it to obtain a textual

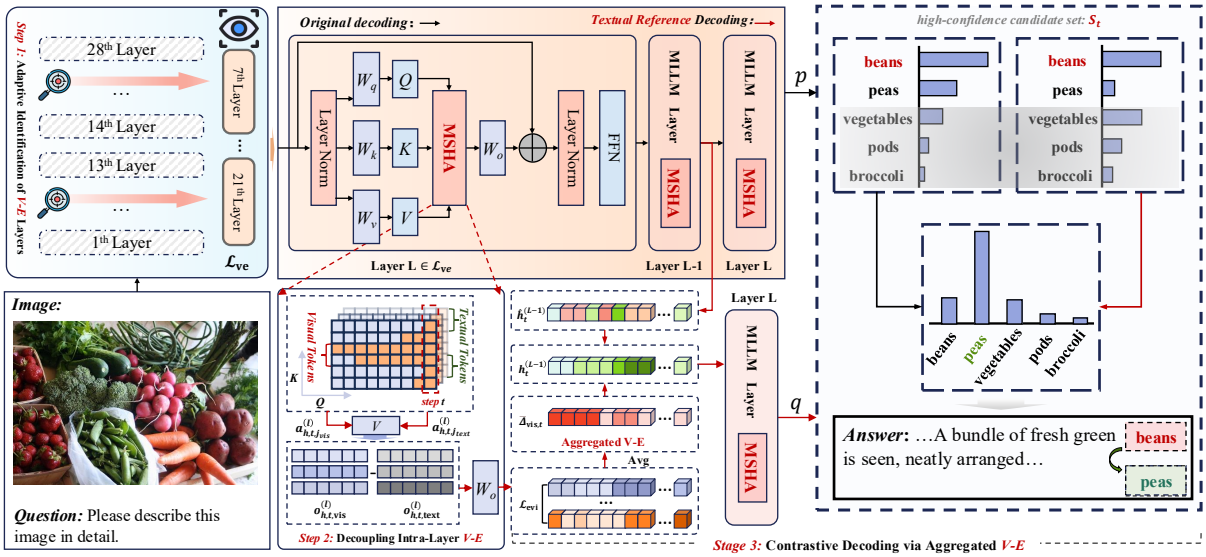


Figure 2: Overview of **DiVE**. The framework operates in three steps: (1) Adaptive Identification of V-E Layers: locating visual-evidence layers via structural constraints and entropy; (2) Decoupling Intra-layer V-E: Decoupling visual evidence from attention mechanisms; and (3) Contrastive Decoding via aggregated V-E: suppressing the aggregated evidence to calibrate logits against language priors.

reference distribution, and performs contrastive decoding against this reference to counter language priors and mitigate hallucinations.

3.1 Adaptive Identification of Visual-Evidence Layers

To ensure the subsequent decoupling is reliable, we first identify a subset of transformer layers where the LVLMs exhibit high dependency on visual inputs. We term this subset the *visual-evidence layers* ($\mathcal{L}_{\text{ve}}$), determined via a two-stage procedure: structural prior constraints followed by entropy-based selection.

Structural Prior Constraints. Consider an LVLM with L transformer layers. For attention-based visual-evidence extraction, layer-wise attention signals are not equally reliable: (i) *Early Layers (Task Recognition).* In early layers, the LVLM is primarily engaged in task-intent recognition, and visual tokens may function as passive attention sinks; as a result, early-layer attention is often less representative of semantic visual evidence. (Fan et al., 2025) (ii) *Late Layers (Language Refinement).* After cross-modal information has been fused, late layers increasingly focus on linguistic refinement and answer generation, with reduced reliance on vision tokens, making final layer attention less suitable for extracting visual evidence. (Fan et al., 2025)

Motivated by the stage-wise layer roles, we define a candidate pool $\mathcal{L}_{\text{cand}}$ for visual-evidence extraction by excluding the early and late $\lceil \kappa L \rceil$ layers:

$$\mathcal{L}_{\text{cand}} = \{ l \in \{1, \dots, L\} \mid \lceil \kappa L \rceil < l \leq L - \lceil \kappa L \rceil \} . \tag{1}$$

Entropy-Based Layer Selection. Within the candidate pool $\mathcal{L}_{\text{cand}}$, we further select layers that demonstrate sharp attention specifically towards visual tokens *during the prefill stage*. For a given head h in layer l, let $\mathbf{a}_{h,l}$ denote the attention weights. Concretely, during prefill we take $\mathbf{a}_{h,l}$ as the attention distribution of the last prompt token (the token whose logits are used to predict the first generated token). We extract $\mathbf{a}_{h,l}$ to the indices of visual tokens $\mathcal{I}_{\text{vis}}$ and renormalize them to form a distribution $\mathbf{p}_{h,l}$. We then propose the *Visual-Language Attention Concentration (V-LAC)* score α_l to quantify visual focus:

$$\alpha_l = \frac{1}{H} \sum_{h=1}^{H} \left(1 - \frac{\mathcal{H}(\mathbf{p}_{h,l})}{\log |\mathcal{I}_{\text{vis}}|} \right) , \tag{2}$$

where $\mathcal{H}(\cdot)$ is the shannon entropy (Shannon, 1948). A higher α_l indicates that the layer focuses intensely on specific visual tokens rather than diffusing attention. We define the final set of visual-evidence layers $\mathcal{L}_{\text{ve}}$ by selecting layers with above-average V-LAC scores within the candidate pool:

$$\mathcal{L}_{\text{ve}} = \{ l \in \mathcal{L}_{\text{cand}} \mid \alpha_l > \mu_{\text{cand}} \} , \tag{3}$$

where μ_{cand} is the mean V-LAC score over $\mathcal{L}_{\text{cand}}$. *This set is computed only once per inference, introducing negligible overhead.*

3.2 Decoupling Intra-layer Visual Evidence

At each decoding step t, for each identified layer $l \in \mathcal{L}_{\text{ve}}$, we extract the intra-layer visual evidence from the attention contribution to the residual stream.

Head-Level Attention Decomposition. Consider the attention mechanism of head h in layer l at decoding step t. The attention weights $a^{(l)}_{h,t,j}$ quantify the attention allocated by the query at step t to the key at position j. To decompose the attention output by modality, we partition the total output $\mathbf{o}^{(l)}_{h,t}$ into two additive vector components based on the token indices $\mathcal{I}_{\text{vis}}$ and $\mathcal{I}_{\text{text}}$:

$$\mathbf{o}^{(l)}_{h,t} = \underbrace{\sum_{j \in \mathcal{I}_{\text{vis}}} a^{(l)}_{h,t,j} \mathbf{v}^{(l)}_{h,j}}_{\mathbf{o}^{(l)}_{h,t,\text{vis}}} + \underbrace{\sum_{j \in \mathcal{I}_{\text{text}}} a^{(l)}_{h,t,j} \mathbf{v}^{(l)}_{h,j}}_{\mathbf{o}^{(l)}_{h,t,\text{text}}}, \quad (4)$$

where $\mathbf{o}^{(l)}_{h,t,\text{vis}}$ and $\mathbf{o}^{(l)}_{h,t,\text{text}}$ represent the attention components aggregated from visual and textual tokens, respectively. In parallel, we calculate the accumulated attention mass for each modality: $m^{(l)}_{h,t,\text{vis}} = \sum_{j \in \mathcal{I}_{\text{vis}}} a^{(l)}_{h,t,j}$ and $m^{(l)}_{h,t,\text{text}} = 1 - m^{(l)}_{h,t,\text{vis}}$.

Decoupling Visual Attention Contribution. However, simply utilizing the raw visual component $\mathbf{o}^{(l)}_{h,t,\text{vis}}$ is insufficient, as it is biased by the probability mass redistribution inherent to Softmax normalization. To decouple the *effective* visual attention contribution from the textual context, we define it as the shift in the output vector induced by visual tokens. To quantify this shift, we first formulate a text-only attention vector derived solely from the textual context, denoted as $\tilde{\mathbf{o}}^{(l)}_{h,t}$. This vector corresponds to the renormalized attention output over the text tokens $\mathcal{I}_{\text{text}}$:

$$\tilde{\mathbf{o}}^{(l)}_{h,t} = \frac{\mathbf{o}^{(l)}_{h,t,\text{text}}}{m^{(l)}_{h,t,\text{text}} + \varepsilon}. \quad (5)$$

The head-wise visual contribution vector at step t, denoted as $\boldsymbol{\delta}^{(l)}_{h,t}$, is then calculated as the deviation of the full attention output from this text-only attention vector:

$$\boldsymbol{\delta}^{(l)}_{h,t} = \mathbf{o}^{(l)}_{h,t} - \tilde{\mathbf{o}}^{(l)}_{h,t}. \quad (6)$$

By substituting $m^{(l)}_{h,t,\text{text}} = 1 - m^{(l)}_{h,t,\text{vis}}$, this yields the *effective* visual contribution:

$$\boldsymbol{\delta}^{(l)}_{h,t} = \mathbf{o}^{(l)}_{h,t,\text{vis}} - \frac{m^{(l)}_{h,t,\text{vis}}}{m^{(l)}_{h,t,\text{text}} + \varepsilon} \mathbf{o}^{(l)}_{h,t,\text{text}}. \quad (7)$$

Eq. 7 shows that $\boldsymbol{\delta}^{(l)}_{h,t}$ equals the raw visual component minus a renormalization-induced correction term, capturing the *effective* visual contribution to the residual stream.

Head-to-Layer Evidence Projection. We obtain a layer-wise visual evidence vector by projecting the concatenated head-wise contributions through the attention output projection matrix $\mathbf{W}^{(l)}_O$:

$$\boldsymbol{\Delta}^{(l)}_{\text{vis},t} = \mathbf{W}^{(l)}_O \left[\boldsymbol{\delta}^{(l)}_{1,t}; \ldots ; \boldsymbol{\delta}^{(l)}_{H,t} \right]. \quad (8)$$

Since $\boldsymbol{\Delta}^{(l)}_{\text{vis},t}$ shares the same vector space as the residual stream (Bietti et al., 2023), it serves as a valid steering vector for calibration in subsequent layers.

3.3 Contrastive Decoding via Aggregated Visual Evidence

Leveraging the decoupled evidence $\boldsymbol{\Delta}^{(l)}_{\text{vis},t}$, we construct a textual reference distribution by suppressing aggregated visual evidence in the penultimate layer, thereby enabling contrastive decoding to counter language priors.

Aggregated Evidence Suppression. Let $\mathbf{h}^{(L-1)}_t$ denote the hidden state of the generated token at decoding step t within the penultimate layer $L-1$. We compute the mean of $\boldsymbol{\Delta}^{(l)}_{\text{vis},t}$ across $l \in \mathcal{L}_{\text{ve}}$ to obtain an aggregated evidence vector $\bar{\boldsymbol{\Delta}}_{\text{vis},t}$:

$$\bar{\boldsymbol{\Delta}}_{\text{vis},t} = \frac{1}{|\mathcal{L}_{\text{ve}}|} \sum_{l \in \mathcal{L}_{\text{ve}}} \boldsymbol{\Delta}^{(l)}_{\text{vis},t}. \quad (9)$$

We directly suppress the aggregated evidence direction $\bar{\boldsymbol{\Delta}}_{\text{vis},t}$ from $\mathbf{h}^{(L-1)}_t$ to obtain $\hat{\mathbf{h}}^{(L-1)}_t$:

$$\hat{\mathbf{h}}^{(L-1)}_t = \mathbf{h}^{(L-1)}_t - \gamma \frac{\|\mathbf{h}^{(L-1)}_t\|_2}{\|\bar{\boldsymbol{\Delta}}_{\text{vis},t}\|_2 + \varepsilon} \bar{\boldsymbol{\Delta}}_{\text{vis},t}, \quad (10)$$

where γ is a hyperparameter controlling suppression strength. We target the penultimate layer for this intervention, as it encodes the high-level semantic representation feeding into the late layer. The modified state propagates through the late layer to the LM head, yielding an evidence-suppressed reference distribution $q_{\text{text}}(y_t \mid y_{<t})$.

Adaptive Likelihood-Ratio Calibration. Given the original LVLM distribution $p_{\text{origin}}(y_t \mid y_{<t})$ and the textual reference $q_{\text{text}}(y_t \mid y_{<t})$, we define a dynamic high-confidence candidate set S_t:

$$S_t = \big\{ y \in \mathcal{V} \;\big|\; p_{\text{origin}}(y \mid y_{<t}) > \tau \cdot \max_{u \in \mathcal{V}} p_{\text{origin}}(u \mid y_{<t}) \big\}. \tag{11}$$

Within this set, we calibrate probabilities using the likelihood ratio between the original distribution and the textual reference, while preserving the total probability mass of S_t:

$$\tilde{p}(y_t \mid y_{<t}) = \mathcal{Z}_t \cdot \frac{p_{\text{origin}}(y_t \mid y_{<t})}{q_{\text{text}}(y_t \mid y_{<t})}, \quad \forall y_t \in S_t, \tag{12}$$

where $\mathcal{Z}_t$ is chosen to preserve the original probability mass of S_t:

$$\mathcal{Z}_t = \frac{\sum_{y \in S_t} p_{\text{origin}}(y \mid y_{<t})}{\sum_{y \in S_t} \big(p_{\text{origin}}(y \mid y_{<t}) / q_{\text{text}}(y \mid y_{<t}) \big)}. \tag{13}$$

Equivalently, for $y_t \in S_t$, this likelihood-ratio reweighting can be expressed in log-space by subtracting $\log q_{\text{text}}$ from $\log p_{\text{origin}}$ and adding a bias term β_t to preserve mass within S_t.

$$y_t \sim \text{softmax}\left[\log p_{\text{origin}}(y_t \mid y_{<t}) - \log q_{\text{text}}(y_t \mid y_{<t}) + \beta_t \right], \tag{14}$$

Here, the likelihood-ratio adjustment is applied only to tokens in S_t; for $y \notin S_t$, we keep the original scores $\log p_{\text{origin}}(y \mid y_{<t})$ to ensure stability. The bias term $\beta_t = \log \mathcal{Z}_t$ preserves the probability mass within S_t.

4 Experiments

We evaluate DiVE against state-of-the-art decoding-based mitigation methods across three standard hallucination benchmarks.

4.1 Experiment Setup

LVLMs and Baselines. We conduct experiments on two representative open-source LVLMs: the widely used LLaVA-v1.5-7B (Liu et al., 2023) and the recently released Qwen2.5-VL-7B (Bai et al., 2025), which stands as the state-of-the-art among 7B-parameter LVLMs. We compare DiVE against five competitive baselines: VCD (Leng et al., 2024), M3ID (Favero et al., 2024), SID (Huo et al., 2025), ONLY (Wan et al., 2025), and AVISC (Woo et al., 2025). To ensure a fair comparison, we utilize the official implementations for all baselines. More details can be found in Appendix A.1

Benchmarks. We utilize three diverse datasets specifically designed to detect and measure visual hallucinations: (1) **POPE:** (Li et al., 2023) A discriminative benchmark that evaluates the reliability of LVLMs in detecting object presence through Yes/No questions under three sampling strategies (Random, Popular, Adversarial). We conduct extensive tests on the MS-COCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019). (2) **CHAIR:** (Rohrbach et al., 2018) A generative benchmark that assesses object hallucinations in image captioning using 500 randomly sampled images from the MS-COCO (Lin et al., 2014) validation set. The evaluation employs three key metrics, CHAIR_S, CHAIR_I and Recall, under max new tokens of 512. (3) **MME-Hallucination:** (Fu et al., 2025) The hallucination subset of MME, consisting of four subsets: existence and count for object-level hallucinations, and position and color for attribute-level hallucinations. More details can be found in Appendix A.2

Implementation Details. We adopt the official prompt templates to facilitate fair comparison. Regarding hyperparameters, unless otherwise specified, the suppression strength is consistently fixed at $\gamma = 0.5$. For structural prior constraints, we define a pruning ratio $\kappa = 0.05$ for all LVLMs, and explicitly exclude the early $\lceil \kappa L \rceil$ and late $\lceil \kappa L \rceil$ layers, where L is the number of transformer layers. The confidence threshold τ is set to 0.1 for POPE, while for CHAIR and MME-Hallucination, it is adjusted to 0.85 to retain only high-confidence candidates. All evaluations are performed on NVIDIA A100 SXM4 (80GB) GPUs.

4.2 Results and Discussions

Results on POPE. As detailed in Tables 1 and 2, DiVE demonstrates robust improvements across benchmarks. On the Qwen2.5-VL-7B backbone, DiVE achieves state-of-the-art performance, surpassing the strong competitor AVISC by +1.63, +1.28, and +1.40 average F1 on MS-COCO, A-OKVQA, and GQA, respectively. Improvements are even more pronounced on LLaVA-v1.5-7B in challenging adversarial subsets; for instance, DiVE outperforms AVISC by +6.36 accuracy on GQA (Adversarial). These results validate DiVE's ability to balance precision and recall effectively.

Table 1: **Performance comparison on POPE using Qwen2.5-VL-7B.** Best results are highlighted in dark green, second-best in light green.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MS-COCO	*base* (Bai et al., 2025)	84.30 ↑0.0	81.42 ↑0.0	83.60 ↑0.0	80.75 ↑0.0	83.23 ↑0.0	80.41 ↑0.0	83.71 ↑0.0	80.86 ↑0.0
	VCD (Leng et al., 2024)	84.70 ↑0.4	81.90 ↑0.5	84.00 ↑0.4	81.35 ↑0.6	83.67 ↑0.4	80.92 ↑0.5	84.12 ↑0.4	81.39 ↑0.5
	M3ID (Favero et al., 2024)	84.57 ↑0.3	81.61 ↑0.2	84.07 ↑0.5	81.21 ↑0.5	83.60 ↑0.4	80.62 ↑0.2	84.08 ↑0.4	81.14 ↑0.3
	SID (Huo et al., 2025)	86.13 ↑1.8	83.92 ↑2.5	85.13 ↑1.5	83.25 ↑2.5	84.47 ↑1.2	82.35 ↑1.9	85.24 ↑1.5	83.17 ↑2.3
	AVISC (Woo et al., 2025)	87.33 ↑3.0	85.70 ↑4.3	86.53 ↑2.9	84.94 ↑4.2	85.40 ↑2.2	83.87 ↑3.5	86.42 ↑2.7	84.84 ↑4.0
	DiVE	88.70 ↑4.4	87.47 ↑6.0	87.77 ↑4.2	86.57 ↑5.8	86.50 ↑3.3	85.38 ↑5.0	87.66 ↑4.0	86.47 ↑5.6
A-OKVQA	*base* (Bai et al., 2025)	88.00 ↑0.0	86.81 ↑0.0	86.03 ↑0.0	84.98 ↑0.0	80.20 ↑0.0	79.96 ↑0.0	84.74 ↑0.0	83.92 ↑0.0
	VCD (Leng et al., 2024)	88.57 ↑0.6	87.30 ↑0.5	86.40 ↑0.4	85.25 ↑0.3	81.73 ↑1.5	81.24 ↑1.3	85.57 ↑0.8	84.60 ↑0.7
	M3ID (Favero et al., 2024)	88.30 ↑0.3	87.97 ↑1.2	86.20 ↑0.2	86.70 ↑1.7	80.43 ↑0.2	80.52 ↑0.6	84.98 ↑0.2	85.06 ↑1.1
	SID (Huo et al., 2025)	87.23 ↓0.8	86.27 ↓0.5	86.53 ↑0.5	85.40 ↑0.4	82.00 ↑1.8	81.26 ↑1.3	85.25 ↑0.5	84.31 ↑0.4
	AVISC (Woo et al., 2025)	89.40 ↑1.4	88.75 ↑1.9	87.33 ↑1.3	86.88 ↑1.9	81.17 ↑1.0	81.93 ↑2.0	85.97 ↑1.2	85.85 ↑1.9
	DiVE	90.93 ↑2.9	90.64 ↑3.8	88.33 ↑2.3	88.27 ↑3.3	81.37 ↑1.2	82.49 ↑2.5	86.88 ↑2.1	87.13 ↑3.2
GQA	*base* (Bai et al., 2025)	87.93 ↑0.0	86.88 ↑0.0	81.23 ↑0.0	80.99 ↑0.0	80.83 ↑0.0	80.66 ↑0.0	83.33 ↑0.0	82.84 ↑0.0
	VCD (Leng et al., 2024)	88.03 ↑0.1	86.64 ↓0.2	84.63 ↑3.4	83.53 ↑2.5	82.03 ↑1.2	81.24 ↑0.6	84.90 ↑1.6	83.80 ↑1.0
	M3ID (Favero et al., 2024)	88.47 ↑0.5	86.92 ↑0.0	82.70 ↑1.5	81.64 ↑0.6	82.27 ↑1.4	81.16 ↑0.5	84.48 ↑1.2	83.24 ↑0.4
	SID (Huo et al., 2025)	88.27 ↑0.3	87.20 ↑0.3	83.27 ↑2.0	82.47 ↑1.5	81.90 ↑1.1	80.94 ↑0.3	84.48 ↑1.2	83.54 ↑0.7
	AVISC (Woo et al., 2025)	88.37 ↑0.4	87.42 ↑0.4	84.03 ↑2.8	83.53 ↑2.5	81.63 ↑0.8	81.50 ↑0.8	84.70 ↑1.4	84.15 ↑1.3
	DiVE	89.97 ↑2.0	89.43 ↑2.6	84.60 ↑3.4	84.64 ↑3.7	82.10 ↑1.3	82.58 ↑1.9	85.56 ↑2.2	85.55 ↑2.7

Table 2: **Performance comparison on POPE using LLaVA-v1.5-7B.**

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MS-COCO	*base* (Liu et al., 2023)	87.09 ↑0.0	86.14 ↑0.0	85.13 ↑0.0	84.28 ↑0.0	79.20 ↑0.0	81.21 ↑0.0	83.81 ↑0.0	83.88 ↑0.0
	VCD (Leng et al., 2024)	87.20 ↑0.1	87.34 ↑1.2	83.47 ↓1.7	84.30 ↑0.0	79.27 ↑0.1	80.23 ↓1.0	83.31 ↓0.5	83.96 ↑0.1
	M3ID (Favero et al., 2024)	86.57 ↓0.5	86.43 ↑0.3	84.27 ↓0.9	84.38 ↑0.1	79.13 ↓0.1	81.52 ↑0.3	83.32 ↓0.5	84.11 ↑0.2
	SID (Huo et al., 2025)	89.10 ↑2.0	89.21 ↑3.1	85.97 ↑0.8	85.29 ↑1.0	80.33 ↑1.1	81.24 ↑0.0	85.13 ↑1.3	85.25 ↑1.4
	ONLY (Wan et al., 2025)	89.73 ↑2.6	89.15 ↑3.0	85.87 ↑0.7	86.25 ↑2.0	79.37 ↑0.2	80.57 ↓0.6	84.99 ↑1.2	85.32 ↑1.4
	AVISC (Woo et al., 2025)	86.63 ↓0.5	84.85 ↓1.3	85.73 ↑0.6	83.68 ↓0.6	82.03 ↑2.8	81.64 ↑0.4	84.80 ↑1.0	83.39 ↓0.5
	DiVE	86.83 ↓0.3	85.13 ↓1.0	85.93 ↑0.8	84.28 ↑0.0	83.83 ↑4.6	82.34 ↑1.1	85.53 ↑1.7	83.92 ↑0.0
A-OKVQA	*base* (Liu et al., 2023)	86.30 ↑0.0	87.49 ↑0.0	80.33 ↑0.0	82.34 ↑0.0	68.33 ↑0.0	75.75 ↑0.0	78.32 ↑0.0	81.86 ↑0.0
	VCD (Leng et al., 2024)	84.17 ↓2.1	86.54 ↓0.9	78.43 ↓1.9	81.32 ↓1.0	68.53 ↑0.2	75.32 ↓0.4	77.04 ↓1.3	81.06 ↓0.8
	M3ID (Favero et al., 2024)	84.53 ↓1.8	86.27 ↓1.2	77.80 ↓2.5	81.21 ↓1.1	69.30 ↑1.0	76.21 ↑0.5	77.21 ↓1.1	81.23 ↓0.6
	SID (Huo et al., 2025)	87.27 ↑1.0	88.43 ↑0.9	81.27 ↑0.9	82.50 ↑0.2	71.23 ↑2.9	77.49 ↑1.7	79.92 ↑1.6	82.81 ↑1.0
	ONLY (Wan et al., 2025)	86.13 ↓0.2	87.18 ↓0.3	79.07 ↓1.3	81.05 ↓1.3	69.03 ↑0.7	76.94 ↑1.2	78.08 ↓0.2	81.72 ↓0.1
	AVISC (Woo et al., 2025)	85.23 ↓1.1	86.33 ↓1.2	82.87 ↑2.5	84.37 ↑2.0	73.37 ↑5.0	77.05 ↑1.3	80.49 ↑2.2	82.58 ↑0.7
	DiVE	90.63 ↑4.3	90.23 ↑2.7	87.43 ↑7.1	87.31 ↑5.0	79.87 ↑11.5	81.11 ↑5.4	85.98 ↑7.7	86.22 ↑4.4
GQA	*base* (Liu et al., 2023)	85.43 ↑0.0	86.75 ↑0.0	79.57 ↑0.0	81.90 ↑0.0	68.70 ↑0.0	74.92 ↑0.0	77.90 ↑0.0	81.19 ↑0.0
	VCD (Leng et al., 2024)	84.17 ↓1.3	85.75 ↓1.0	74.10 ↓5.5	76.43 ↓5.5	69.13 ↑0.4	75.24 ↑0.3	75.80 ↓2.1	79.14 ↓2.0
	M3ID (Favero et al., 2024)	84.57 ↓0.9	86.20 ↓0.5	74.17 ↓5.4	78.35 ↓3.6	69.23 ↑0.5	76.37 ↑1.5	75.99 ↓1.9	80.31 ↓0.9
	SID (Huo et al., 2025)	85.37 ↓0.1	88.15 ↑1.4	74.97 ↓4.6	80.12 ↓1.8	69.57 ↑0.9	76.30 ↑1.4	76.64 ↓1.3	81.52 ↑0.3
	ONLY (Wan et al., 2025)	85.97 ↑0.5	88.10 ↑1.3	74.27 ↓5.3	78.60 ↓3.3	69.57 ↑0.9	75.87 ↑1.0	76.60 ↓1.3	80.86 ↓0.3
	AVISC (Woo et al., 2025)	87.43 ↑2.0	88.25 ↑1.5	79.43 ↓0.1	82.53 ↑0.6	74.67 ↑6.0	78.13 ↑3.2	80.51 ↑2.6	82.97 ↑1.8
	DiVE	88.83 ↑3.4	88.32 ↑1.6	85.00 ↑5.4	84.92 ↑3.0	81.03 ↑12.3	81.66 ↑6.7	84.95 ↑7.0	84.97 ↑3.8

Table 3: **Results on MME-Hallucination.** Comparison of object and attribute-level scores.

Model	Method	Object-level		Attribute-level		Total Score
		Existence	Count	Position	Color	
Qwen2.5-VL-7B	*base* (Bai et al., 2025)	195.0 ↑0.0	165.0 ↑0.0	140.0 ↑0.0	190.0 ↑0.0	690.0 ↑0.0
	VCD (Leng et al., 2024)	195.0 ↑0.0	165.0 ↑0.0	145.0 ↑5.0	190.0 ↑0.0	695.0 ↑5.0
	M3ID (Favero et al., 2024)	195.0 ↑0.0	165.0 ↑0.0	150.0 ↑10.0	190.0 ↑0.0	700.0 ↑10.0
	SID (Huo et al., 2025)	195.0 ↑0.0	153.3 ↓11.7	165.0 ↑25.0	190.0 ↑0.0	703.3 ↑13.3
	AVISC (Woo et al., 2025)	195.0 ↑0.0	153.3 ↓11.7	165.0 ↑25.0	195.0 ↑5.0	708.3 ↑18.3
	DiVE	195.0 ↑0.0	165.0 ↑0.0	165.0 ↑25.0	195.0 ↑5.0	720.0 ↑30.0
LLaVA-v1.5-7B	*base* (Liu et al., 2023)	190.0 ↑0.0	155.0 ↑0.0	128.3 ↑0.0	170.0 ↑0.0	643.3 ↑0.0
	VCD (Leng et al., 2024)	190.0 ↑0.0	155.0 ↑0.0	133.3 ↑5.0	170.0 ↑0.0	648.3 ↑5.0
	M3ID (Favero et al., 2024)	195.0 ↑5.0	160.0 ↑5.0	133.3 ↑5.0	165.0 ↓5.0	653.3 ↑10.0
	SID (Huo et al., 2025)	190.0 ↑0.0	158.3 ↑3.3	136.7 ↑8.4	165.0 ↓5.0	650.0 ↑6.7
	ONLY (Wan et al., 2025)	195.0 ↑5.0	160.0 ↑5.0	133.3 ↑5.0	165.0 ↓5.0	653.3 ↑10.0
	AVISC (Woo et al., 2025)	195.0 ↑5.0	156.7 ↑1.7	143.3 ↑15.0	170.0 ↑0.0	665.0 ↑21.7
	DiVE	195.0 ↑5.0	163.3 ↑8.3	143.3 ↑15.0	170.0 ↑0.0	671.7 ↑28.4

Results on MME-Hallucination. We further extend evaluation to the MME-Hallucination, with detailed results provided in Table 3. DiVE achieves the highest total scores across both backbones. With Qwen2.5-VL-7B, DiVE reaches 720.0, surpassing second-best AVISC by 11.7 points, mainly due to a stronger Object-level *Count* score. Compared to the base model, DiVE yields a clear attribute-level gain, notably improving *Position* by 25.0 points (140.0 → 165.0). Similarly, on LLaVA-v1.5-7B, DiVE attains 671.7, exceeding VCD by 23.4 points (and outperforming AVISC by 6.7 points), demonstrating consistent benefits on both Object-level and Attribute-level scores.

Results on CHAIR. Table 4 further demonstrates the effectiveness of DiVE on open-ended generation. Across both backbones, DiVE achieves the best CHAIR scores among compared methods, indicating consistently reduced hallucination rates. On Qwen2.5-VL-7B, DiVE lowers CHAIR_S to 31.4 (2.4 points below the second-best M3ID) and CHAIR_I to 7.3 (1.6 points below AVISC), while also attaining the highest Recall of 73.0. On LLaVA-v1.5-7B, DiVE reaches 47.4, 13.0 on CHAIR_S and CHAIR_I, improving over the strong SID baseline by 2.8 points in sentence-level hallucinations, while maintaining competitive Recall (81.5, second-best). These results suggest more faithful and accurate image descriptions with max new tokens = 512.

Table 4: **Results on CHAIR.** We report metric scores (↓) and Recall (↑) under a max new tokens of 512.

Model	Method	CHAIR_S ↓	CHAIR_I ↓	Recall ↑
Qwen2.5-VL-7B	*base* (Bai et al., 2025)	37.6 ↑0.0	9.3 ↑0.0	71.7 ↑0.0
	VCD (Leng et al., 2024)	37.8 ↑0.2	9.4 ↑0.1	70.8 ↓0.9
	M3ID (Favero et al., 2024)	33.8 ↓3.8	9.2 ↓0.1	69.3 ↓2.4
	SID (Huo et al., 2025)	36.2 ↓1.4	9.1 ↓0.2	72.5 ↑0.8
	AVISC (Woo et al., 2025)	35.4 ↓2.2	8.9 ↓0.4	70.7 ↓1.0
	DiVE	31.4 ↓6.2	7.3 ↓2.0	73.0 ↑1.3
LLaVA-v1.5-7B	*base* (Liu et al., 2023)	53.2 ↑0.0	14.7 ↑0.0	80.5 ↑0.0
	VCD (Leng et al., 2024)	51.2 ↓2.0	14.1 ↓0.6	79.7 ↓0.8
	M3ID (Favero et al., 2024)	54.6 ↑1.4	14.8 ↑0.1	82.0 ↑1.5
	SID (Huo et al., 2025)	50.2 ↓3.0	13.3 ↓1.4	78.9 ↓1.6
	ONLY (Wan et al., 2025)	50.8 ↓2.4	14.5 ↓0.2	80.2 ↓0.3
	AVISC (Woo et al., 2025)	51.6 ↓1.6	13.8 ↓0.9	79.3 ↓1.2
	DiVE	47.4 ↓5.8	13.0 ↓1.7	81.5 ↑1.0

4.3 Ablation Study

Effectiveness of Structural Prior Constraints. Figure 3 highlights the importance of our layer constraints. DiVE achieves the lowest hallucination levels on both LVLMs, outperforming @Early and @Late. On Qwen2.5-VL-7B, DiVE reduces CHAIR_S and CHAIR_I from 37.6, 9.3 to 31.4, 7.3, whereas retaining only early or late layers yields weaker results (34.4, 8.9 and 33.8, 8.3). This suggests early-layer attention can be distorted by sink-like behavior, while late-layer attention is more influenced by language priors during answer generation. *The similar trend holds on LLaVA-v1.5-7B*. Overall, these results support excluding early and late layers when constructing the candidate pool for visual-evidence extraction.

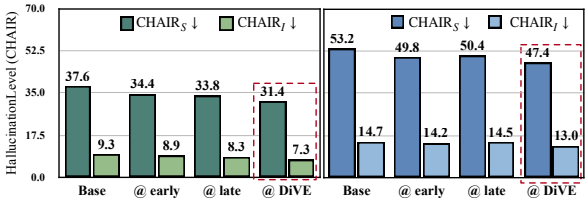


Figure 3: **Validation of Structural Prior Constraints.** We compare retaining the specific region in the candidate pool (@Early (initial 5%) and @Late (final 5%) vs. DiVE on Qwen2.5-VL-7B (left) and LLaVA-v1.5-7B (right).

Table 5: **Impact of exclusion ratio κ.** Comparison of performance with varying κ on Qwen2.5-VL-7B and LLaVA-v1.5-7B.

Model	Values	POPE		CHAIR		MME
		Accuracy ↑	F1 ↑	CHAIR_S ↓	CHAIR_I ↓	
Qwen2.5-VL-7B	*base*	83.93	82.54	37.6	9.3	690.0
	$\kappa = 0.00$	85.31	84.61	33.8	7.6	703.3
	$\kappa = 0.02$	86.48	86.06	32.0	8.1	703.3
	$\kappa = \mathbf{0.05}$	86.70	86.38	31.4	7.3	720.0
	$\kappa = 0.08$	86.62	86.24	32.2	7.4	713.3
	$\kappa = 0.11$	86.59	86.18	32.6	8.1	703.3
LLaVA-v1.5-7B	*base*	80.01	82.31	53.2	14.7	643.3
	$\kappa = 0.00$	84.12	83.23	50.6	14.9	651.7
	$\kappa = 0.02$	84.50	83.60	48.4	14.2	651.7
	$\kappa = \mathbf{0.05}$	85.49	85.03	47.4	13.0	671.7
	$\kappa = 0.08$	84.47	83.97	49.0	14.0	656.7
	$\kappa = 0.11$	83.87	83.14	51.4	15.0	656.7

Impact of Exclusion Ratio κ. We investigate the sensitivity of our approach to the exclusion ratio κ, which drops the early/late $\lceil \kappa L \rceil$ layers (with L being the total number of layers). As evidenced in Table 5, performance consistently peaks at $\kappa = 0.05$ across all metrics on both backbones, yielding the lowest hallucination rates (e.g., CHAIR_S of 31.4 and 47.4) and the highest MME score (720.0 and 671.7). Lower ratios ($\kappa < 0.05$) may be insufficient to fully exclude noisy regions, whereas excessive pruning ($\kappa > 0.05$) degrades performance (e.g., MME drops to 703.3 on Qwen2.5-VL-7B and 656.7 on LLaVA-v1.5-7B at $\kappa = 0.11$). Therefore, we adopt $\kappa = 0.05$ as the default trade-off between noise reduction and evidence preservation.

Table 6: **Impact of suppression strength γ.** Comparison of performance with varying γ on Qwen2.5-VL-7B and LLaVA-v1.5-7B.

Model	Values	POPE		CHAIR		MME-Hallu
		Accuracy ↑	F1 ↑	$CHAIR_S$ ↓	$CHAIR_I$ ↓	
Qwen2.5-VL-7B	*base*	83.93	82.54	37.6	9.3	690.0
	$\gamma = 0.10$	85.86	84.85	34.0	7.7	708.3
	$\gamma = 0.30$	86.28	85.87	32.8	7.4	708.3
	γ **= 0.50**	86.70	86.38	31.4	7.3	720.0
	$\gamma = 0.80$	86.13	85.02	32.6	7.3	713.3
	$\gamma = 1.00$	85.90	85.01	34.2	7.6	713.3
LLaVA-v1.5-7B	*base*	80.01	82.31	53.2	14.7	643.3
	$\gamma = 0.10$	85.12	84.79	51.6	14.7	650.0
	$\gamma = 0.30$	85.33	84.87	49.4	14.3	656.7
	γ **= 0.50**	85.49	85.03	47.4	13.0	671.7
	$\gamma = 0.80$	84.97	84.34	49.8	14.0	671.7
	$\gamma = 1.00$	84.87	83.21	50.4	14.5	671.7

Impact of Visual Suppression Strength γ. We analyze the sensitivity of our method to the suppression strength γ (Eq. 10), which determines the intensity of the visual steering vector subtracted from the hidden state. As shown in Table 6, performance consistently peaks at $\gamma = 0.5$. A moderate value of $\gamma = 0.5$ effectively suppress the visual evidence to construct a reliable textual reference. Lower values ($\gamma < 0.5$) result in insufficient suppression, where the reference representation remains entangled with visual information, reducing the contrastive effect. Conversely, excessive suppression ($\gamma > 0.5$) overly alters the hidden states, distorting the semantic features and leading to performance degradation. Consequently, we adopt $\gamma = 0.5$ as the optimal setting to balance evidence removal and feature stability. *More parameter analysis are provided in Appendix B.1.*

Table 7: **Ablation on layer selection strategy.** Comparison of different selection metrics on Qwen2.5-VL-7B.

Strategy	POPE		CHAIR		MME-Hallu ↑
	Accuracy ↑	F1 ↑	$CHAIR_S$ ↓	$CHAIR_I$ ↓	
base	83.93	82.54	37.6	9.3	690.0
$a_l \rightarrow r_l$	84.17	83.26	36.2	8.6	708.3
$a_l \rightarrow r_l \times a_l$	85.60	84.85	33.6	7.8	710.0
$\mathcal{L}_{\text{evi}} \rightarrow \{\arg\max \alpha_l\}$	84.58	83.64	35.2	8.1	688.3
DiVE	86.70	86.38	31.4	7.3	720.0

Analysis of Layer Selection Strategy. Table 7 validates the necessity of our V-LAC criterion (α_l). We define the visual attention ratio r_l:

$$r_l = \frac{1}{H} \sum_{h=1}^{H} \frac{\sum_{i \in \mathcal{I}_{\text{vis}}} A_{h,l}(i)}{\sum_i A_{h,l}(i)}. \tag{15}$$

The $A_{h,l}$ denote the attention distribution of the token used to predict the first generated token. Replacing a_l with the visual attention ratio r_l markedly worsens performance ($CHAIR_S$: 31.4 → 36.2), indicating that the visual attention mass ratio alone is not a reliable proxy for effective visual evidence. Weighting a_l by r_l ($r_l \times a_l$) remains suboptimal ($CHAIR_S$ = 33.6), suggesting that evidence quality captured by a_l is more informative than simply attending more. Moreover, selecting only the single best layer ($\arg\max a_l$) collapses performance (MME-Hallu: 720.0 → 688.3), implying that visual evidence is distributed across multiple layers, and our dynamic thresholding better preserves this multi-layer evidence.

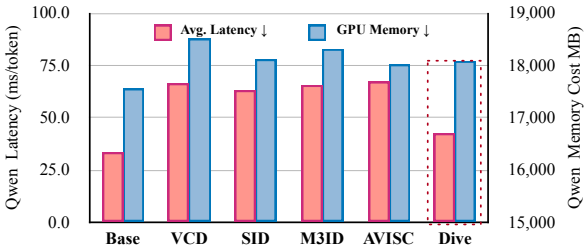


Figure 4: **Inference efficiency on Qwen-2.5-VL-7B.** We report the average per-token latency (ms) and peak GPU memory cost (MB).

4.4 Efficiency Comparison

As illustrated in Figure 4, conventional contrastive methods (e.g., VCD, M3ID) incur a substantial latency penalty ($\sim 2\times$) due to dual forward passes. In contrast, DiVE maintains inference speeds comparable to the Base model by restricting intervention to a sparse subset of evidence layers. This ensures efficient hallucination mitigation with negligible memory overhead. *Detailed numerical results are provided in Appendix B.2.*

5 Conclusion

In this work, we present **DiVE**, a training-free decoding framework to mitigate visual hallucinations in LVLMs. Addressing the limitations of prior internal methods in decoupling effective visual evidence from textual context, DiVE implements adaptive *visual-evidence layer* identification and intra-layer decoupling. This process extract aggregated visual evidence directly from attention mechanisms. By suppressing this evidence, we construct a reliable textual reference distribution to contrastive decoding against language priors. Extensive evaluations confirm that DiVE achieves state-of-the-art results on multiple hallucination benchmarks. Crucially, DiVE operates without the latency of extra forward passes or external perturbations, offering a highly adaptable and efficient solution for LVLMs' visual hallucination mitigation.

6 Limitations

DiVE effectively mitigates hallucinations but remains challenged in scenarios demanding fine-grained perception and precise spatial grounding (Figure 6). When visual evidence is weak or ambiguous, the LVLMs can still confuse visually similar categories, misidentify small or partially occluded objects, or accept incorrect spatial premises. Dense contexts also amplify object co-occurrence biases, where plausible associations (e.g., desk→bookshelf, street→traffic lights) can still compete with subtle visual cues. DiVE introduces additional test-time computation due to intra-layer evidence extraction and the textual-reference construction, but the practical overhead is modest: 42.6 ms/token ($1.28\times$) with negligible memory increase ($1.03\times$), substantially more efficient than contrastive baselines that nearly double latency. Still, DiVE consistently mitigate hallucination and yields significant gains.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. 2023. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36:1560–1588.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Representation Learning*, volume 2024, pages 54158–54183.

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.

Yingqi Fan, Anhao Zhao, Jinlan Fu, Junlong Tong, Hui Su, Yijie Pan, Wei Zhang, and Xiaoyu Shen. 2025. Visipruner: Decoding discontinuous cross-modal dynamics for efficient multimodal llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18896–18913.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312. IEEE Computer Society.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

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.

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

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

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.

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

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

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. 2024. Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8228–8248.

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.

Claude 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, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. 2024. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pages 198–215. Springer.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. 2025. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3225–3234.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *ACL (Findings)*.

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.

Jiulong Wu, Zhengliang Shi, Shuaiqiang Wang, Jizhou Huang, Dawei Yin, Lingyong Yan, Min Cao, and Min Zhang. 2025. Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19456–19472. Association for Computational Linguistics.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273. Association for Computational Linguistics.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 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 (CVPR)*, pages 13807–13816.

Xinping Zhao, Jindi Yu, Zhenyu Liu, Jifang Wang, Dongfang Li, Yibin Chen, Baotian Hu, and Min Zhang. 2024. Medico: Towards hallucination detection and correction with multi-source evidence fusion. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 34–45.

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. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Appendix

Our appendix is organized as follows:

- **Section A**: Detailed experimental setups, including benchmark definitions and baseline formulations.
- **Section B**: Extended ablation studies, including parameter sensitivity and computational cost analysis.
- **Section C**: Comprehensive qualitative comparisons and error analysis.

A More Experimental Details

A.1 Baselines

We compare our approach against the following *state-of-the-art* hallucination mitigation strategies. We provide a brief formulation for each method below.

VCD (Visual Contrastive Decoding). VCD (Leng et al., 2024) mitigates statistical biases and language priors by contrasting the logits derived from the original visual input v against those from a distorted visual input v'. The distorted image v' is typically synthesized by applying a Gaussian noise mask following a forward diffusion process. The calibrated probability distribution is defined as:

$$p_{\text{vcd}}(y \mid v, v', x) = \text{softmax}\Big((1+\alpha)\ \text{logit}(y \mid v, x) - \alpha\ \text{logit}(y \mid v', x)\Big) \tag{16}$$

where α is a hyperparameter controlling the penalty strength. This mechanism enforces the model to favor tokens that are robust to visual perturbations, thereby suppressing hallucinations driven by language priors.

M3ID (Multi-Modal Mutual Information Decoding). M3ID (Favero et al., 2024) mitigates hallucinations by contrasting the token distributions with the image and without the image (text-only prior) during decoding, and adding a corrective term when the model is not confident. It measures the discrepancy between the conditional logits ℓ_c (with image) and the unconditional logits ℓ_u (text-only prior):

$$\ell_c(y_t) = \ell(y_t \mid y_{<t}, x, c), \quad \ell_u(y_t) = \ell(y_t \mid y_{<t}, x). \tag{17}$$

M3ID uses an adaptive weight γ_t (with a confidence gate and a time-dependent schedule) to scale the corrective term $(\ell_c - \ell_u)$.

$$\ell_{\text{final}}(y_t) = \ell_c(y_t) + \gamma_t\big(\ell_c(y_t) - \ell_u(y_t)\big), \tag{18}$$

This adaptive re-weighting amplifies image-grounded signals specifically when the language prior dominates the generation process.

SID (Self-Introspective Decoding). Unlike holistic image distortion methods, SID (Huo et al., 2025) employs a token-level intervention strategy. It utilizes a Context and Text-aware Token Selection (CT^2S) mechanism to compute visual token importance scores based on the attention matrix $A^{(l)}$ from a selected decoder layer l (typically early layers):

$$\text{Score}_{l,i}(v) = \frac{1}{h}\sum_{j=1}^{h} A^{(l)}_{j,-1,i}. \tag{19}$$

SID constructs a "hallucination-inducing" context by retaining only the *least* important visual tokens (those with the lowest scores). The final decoding subtracts the logits generated from this perturbed context from the original logits, thereby effectively filtering out spurious vision-text associations.

AVISC (Attentional Vision Calibration). AVISC (Woo et al., 2025) proposes a training-free calibration method that operates on the attention mechanism. It identifies blind tokens via the image-wise attention distribution and performs a training-free logit calibration by contrasting the original predictions against a blind-token baseline branch. The calibrated probability is approximated as:

$$p_{\text{avisc}}(y_t) \approx \text{softmax}\Big((1+\alpha)\,\ell(y_t \mid v, x) - \alpha\,\ell(y_t \mid v_{\text{blind}}, x)\Big) \tag{20}$$

By down-weighting tokens that remain highly probable under the blind-token baseline branch, AViSC reduces the model's reliance on parametric knowledge and unimodal biases.

ONLY (One-Layer Intervention). ONLY (Wan et al., 2025) posits that effective hallucination mitigation can be achieved through intervention at a single Transformer layer. It introduces the Text-to-Visual Entropy Ratio (TVER) to identify attention heads that disproportionately prioritize textual information:

$$\mathcal{H}_{\text{high}} = \left\{ h : \text{TVER}_{l,h} \geq \frac{1}{H} \sum_{h'=1}^{H} \text{TVER}_{l,h'} \right\}. \tag{21}$$

The method inserts a textual-enhanced multi-head attention (TE-MHA) at a single selected layer to obtain textual-enhanced logits, and then adaptively performs contrastive and collaborative decoding with the original logits.

A.2 Benchmark

To comprehensively assess the extent of hallucination in LVLMs, we conduct experiments across three widely-adopted benchmarks. The detailed configurations and metric definitions are elaborated below.

POPE. (Polling-based Object Probing Evaluation) (Li et al., 2023) serves as a primary metric for assessing object-level hallucinations. It operates by querying the model with binary "Yes-or-No" questions regarding the existence of specific objects in an image (e.g., "*Is there a {object} in the image?*"). To ensure a fair comparison, we strictly employ the standard query format without appending auxiliary prompt (e.g., "*Please answer yes or no*"). The evaluation corpus is constructed based on images sampled from standard datasets, including MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022) and GQA (Hudson and Manning, 2019), each combination of dataset and query type (random, popular, adversarial) consists of 3,000 samples. This probing strategy effectively isolates the model's tendency to hallucinate non-existent objects without the interference of complex language generation metrics.

CHAIR. (Caption Hallucination Assessment with Image Relevance) (Rohrbach et al., 2018) evaluates hallucinations in the context of free-form image captioning. In our experiments, LVLMs are prompted to generate descriptions for 500 randomly selected images from the MSCOCO validation set. For each image, captions are generated by prompting the LVLMs with the prompt: "Please describe this image in detail.". The evaluation metrics categorize hallucinations into two granularities: instance-level (CHAIR_I), sentence-level (CHAIR_S) and recall. These are formally defined as:

$$\begin{aligned}
\text{CHAIR}_\text{I} &= \frac{|\ \{\text{hallucinated objects mentioned}\}\ |}{|\ \{\text{all objects mentioned}\}\ |} \\
\text{CHAIR}_\text{S} &= \frac{|\ \{\text{hallucinated captions}\}\ |}{|\ \{\text{all captions}\}\ |} \\
\text{Recall} &= \frac{|\ \{\text{correctly predicted objects}\}\ |}{|\ \{\text{total ground-truth objects}\}\ |}
\end{aligned} \tag{22}$$

A lower score on these metrics indicates a reduced frequency of hallucinated content and higher generation fidelity.

MME-Hallucination. We further include MME-Hallucination (Fu et al., 2025) for a more granular analysis. This benchmark comprises four distinct subsets targeting different hallucination types: *Existence* and *Count* (object-level), alongside *Position* and *Color* (attribute-level). Each subset contains 30 images paired with 60 binary questions (two per image). Consistent with the protocol in POPE (Li et al., 2023), this benchmark also relies on binary queries. Following the official implementation (Fu et al., 2025), performance is reported using the summation of *Accuracy* and *Accuracy+*. Here, *Accuracy* denotes the correctness of individual answers, while *Accuracy+* acts as a stricter metric, requiring both questions associated with a single image to be answered correctly to count as a pass.

A.3 Dataset and Model Licensing

Datasets. We use established public benchmarks and follow their original access conditions and licenses. POPE (Li et al., 2023) is released under the MIT license. CHAIR (Rohrbach et al., 2018) evaluation code is released under the BSD 2-Clause license. For MS-COCO (Lin et al., 2014), the images remain under their original Flickr licenses, while COCO annotations are distributed under CC BY 4.0. A-OKVQA (Schwenk et al., 2022) is built on MS-COCO images; we follow MS-COCO image terms and use A-OKVQA according to its release conditions (the official repository is Apache-2.0). GQA (Hudson and Manning, 2019)is used under its public release conditions (commonly distributed under CC BY 4.0 in widely used distributions). MME-Hallucination is used strictly for academic research purposes following its benchmark release conditions (commonly described as research-only).

Code and Models. We conduct experiments using open-source LVLM backbones and official implementations when available. LLaVA-v1.5-7B (Liu et al., 2023) is licensed under Apache-2.0.

Table 8: Impact of confidence threshold τ on our model (employ **DiVE** on **Qwen2.5-VL-7B**)

POPE			CHAIR / MME			
τ	Accuracy ↑	F1-score ↑	τ	CHAIR_S ↓	CHAIR_I ↓	MME-Hallu ↑
base	83.93	82.54	*base*	37.6	9.3	690.0
0.10	86.70	86.38	0.70	33.0	7.7	720.0
0.20	86.38	85.87	0.75	32.8	7.6	720.0
0.30	85.71	85.43	0.70	33.4	7.7	720.0
0.40	85.83	85.78	**0.85**	31.4	7.3	720.0
0.50	85.91	85.95	0.90	32.6	7.5	715.0

Table 9: Impact of confidence threshold τ on our model (employ **DiVE** on **LLaVA-v1.5-7B**).

POPE			CHAIR / MME			
τ	Accuracy ↑	F1-score ↑	τ	CHAIR_S ↓	CHAIR_I ↓	MME ↑
base	80.01	82.31	*base*	53.2	14.7	643.3
0.10	85.49	85.03	0.70	50.6	13.8	661.7
0.20	85.23	84.72	0.75	48.4	13.4	661.7
0.30	85.03	84.51	0.80	48.0	13.4	661.7
0.40	85.10	84.62	**0.85**	46.8	13.0	671.7
0.50	84.97	84.26	0.90	49.2	13.7	661.7

Table 10: **Computational cost analysis.** Measurements are averaged over 3 runs on an NVIDIA A100 GPU. Values in gray denote the relative overhead factor (×) compared to the **Qwen2.5-VL-7B**.

Method	GPU Memory (MB) ↓		Avg. Latency (ms/token) ↓	
Base	17568	(×1.00)	33.4	(×1.00)
VCD (Leng et al., 2024)	18513	(×1.05)	65.9	(×1.97)
SID (Huo et al., 2025)	18118	(×1.03)	68.4	(×2.05)
M3ID (Favero et al., 2024)	18309	(×1.04)	65.1	(×1.95)
AVISC (Woo et al., 2025)	17999	(×1.02)	67.1	(×2.01)
DiVE	18078	(×1.03)	42.6	(×1.28)

Qwen2.5-VL-7B (Bai et al., 2025) is licensed under Apache-2.0.

B More Ablation Studies and Analysis

B.1 Sensitivity to Confidence Threshold

We analyze the impact of the confidence threshold τ, which controls the candidate selection scope. As shown in Tables 8 and 9, the optimal threshold varies based on task characteristics. For binary existence probing (POPE), performance peaks at a lower threshold ($\tau = 0.1$), indicating that retaining a broader candidate pool effectively safeguards against false negatives without introducing excessive noise. Conversely, complex generation and attribute reasoning tasks (CHAIR, MME) benefit from stricter constraints ($\tau = 0.85$). In these scenarios, a higher threshold is necessary to rigorously filter out low-confidence candidates driven by language priors, ensuring high-fidelity generation. Consequently, we adopt $\tau = 0.1$ for POPE and $\tau = 0.85$ for CHAIR and MME to achieve optimal performance across diverse benchmarks.

B.2 Detailed Computational Costs

To assess inference efficiency, we report fine-grained measurements of per-token latency and memory consumption in Table 10. Experiments are conducted on an NVIDIA A100 SXM4 (80GB) GPU. Following the standard CHAIR benchmark protocol, we randomly sample 500 instances and set the maximum new token limit to 512. To minimize variance, each sample is decoded three times, and the results are averaged. Quantitative results reveal that while baselines like VCD and M3ID essentially double the inference time (1.97× and 2.05×, respectively), DiVE maintains a streamlined execution flow with an average latency of 42.6 ms/token (only 1.28× overhead). Furthermore, the memory cost increase is negligible (1.03×), validating DiVE as a highly efficient solution for practical deployment.

C More Case Studies and Error Analysis

C.1 Qualitative Comparison

To demonstrate the effectiveness of DiVE across different architectures, we compare qualitative results in Figure 5, where the Top Panel displays

Qwen2.5-VL-7B and the Bottom Panel displays LLaVA-v1.5-7B.

Mitigating Co-occurrence and False Negatives (POPE). DiVE shows robust performance in both rejecting hallucinations and recovering missing details. In the bus depot scenario (Bottom Panel), baselines incorrectly confirm a "motorcycle" due to its co-occurrence with "bus," which DiVE correctly rejects. Conversely, in the kitchen scene (Top Panel), the baseline fails to detect a "spoon," whereas DiVE successfully retrieves this missing object, effectively decoupling visual evidence from statistical priors.

Fine-grained Attribute Grounding (MME). Standard decoding methods often fail to align numerical attributes with visual facts. As shown in the Bottom Panel, baselines fail to distinguish the specific count of "two bottles," while in the Top Panel, they incorrectly affirm "three people." In both cases, DiVE accurately identifies the discrepancies, demonstrating superior sensitivity to fine-grained visual details over holistic sentence probability.

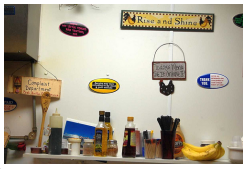
Suppressing Contextual Hallucination (CHAIR). In open-ended captioning, baselines succumb to strong context priors. For instance, in the choir performance (Bottom Panel), baselines hallucinate an "audience" simply because of the "stage" context. DiVE resists this generative hallucination, generating a faithful description focused solely on the visible performers and screen.

C.2 Error Analysis

We further analyze failure cases in Figure 6 to identify common challenges shared across architectures. A primary source of error is *visual uncertainty*, where models struggle with fine-grained discrimination; for instance, Qwen-2.5-VL-7B (Top Panel) confuses a cow with "sheep" due to feature similarity, while LLaVA-v1.5-7B (Bottom Panel) misidentifies a red object in the snow as a "traffic light." *Spatial reasoning* also remains a bottleneck, as seen when Qwen fails to determine the spatial relationship of elephants and LLaVA accepts the incorrect premise of a "light under the computer." Furthermore, *strong object co-occurrence biases* can still override visual evidence in dense contexts: Qwen-2.5-VL-7B hallucinates a "bookshelf" simply due to the presence of a desk, and LLaVA generates "traffic lights" in a busy street scene despite their absence. These results highlight that while DiVE effectively mitigates hallucinations, distinguishing ambiguous visual evidence from strong language priors remains a persistent challenge for the field.

POPE

Visual Input



Question: Is there a spoon in the image?
Label: Yes
Base (Qwen 2.5-VL-7B): No, there is no spoon visible in the image.
VCD: No, there are no spoons visible in the image.
SID: There are no spoons visible in the image.
AVISC: There doesn't appear to be a spoon in the image.
DiVE: Yes, there is a wooden spoon hanging on the left side of the image.

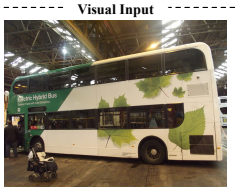
MME-Hallucination

Visual Input

Question: Are there only three people appear in this image? Please answer yes or no.
Label: No
Base (Qwen 2.5-VL-7B): Yes.
VCD: Yes.
SID: Yes.
AVISC: Yes.
DiVE: No.

CHAIR

Visual Input



Question: Please describe this image in detail.
Ground truth words: [“person”, “bus”]
Base (Qwen 2.5-VL-7B): The image shows a double-decker bus parked indoors, likely at a bus station or depot. **...** The interior appears to be visible through the windows, showing passengers seated inside.\n\nIn the foreground, there is a stroller parked near the bus, suggesting the presence of people who might be traveling with children. **…**
VCD: The image shows a double-decker bus parked indoors, likely in a large hangar or garage. ... The interior is visible through the windows, showing passengers seated inside. ... In the foreground, a stroller is parked near the bus, suggesting the presence of a pedestrian, possibly a parent or guardian, who is not fully visible in the frame. ..."

SID: The image shows a double-decker bus stationary inside a large indoor space. ... The lower deck windows are transparent, and it appears that there are passengers seated within the vehicle. ... A stroller is placed near the front, indicating that a pedestrian or a family is likely nearby. ..."
AVISC: The image depicts a double-decker bus parked indoors, possibly at a maintenance depot. ... Through the glass, one can observe figures that look like passengers seated inside. ... and a few people or staff might be in the background, while a stroller sits prominently in the foreground. ..."
DiVE: The image shows a double-decker bus parked indoors, likely in a large hangar or garage. The bus is labeled as an \"Electric Hybrid Bus\" and is designed to promote environmentally friendly travel with the tagline \"Greener Travel with Low Emissions.\" The bus has a predominantly white body with green accents, including large green leaf designs on the side, symbolizing its eco-friendly nature. The text on the side of the bus is in a clean, modern font, and the bus appears to be a modern model, possibly for public transportation.\n\nIn the foreground, there is a stroller with a person sitting in it, suggesting the presence of a child. The stroller is positioned close to the bus, and the person appears to be waiting or observing the bus. The indoor setting has a high ceiling with visible structural beams and lighting fixtures, indicating that the bus is in a maintenance or exhibition area. The floor is made of concrete, and the overall environment looks clean and organized.

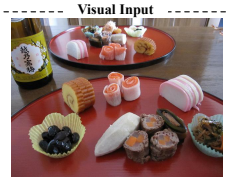
POPE

Visual Input

Question: Is there a motorcycle in the image?
Label: No
Base (LLaVA-1.5-7B): Yes
VCD: Yes
SID: Yes
AVISC: Yes
DiVE: No

MME-Hallucination

Visual Input



Question: Is there two bottles in the image? Please answer yes or no.
Label: No
Base (LLaVA-1.5-7B): Yes.
VCD: Yes.
SID: Yes.
AVISC: Yes.
DiVE: No.

CHAIR

Visual Input

Question: Please describe this image in detail.
Ground truth words: [“person”, “tie”]
Base (LLaVA 1.5-VL-7B): The image features a group of young men standing on a stage, dressed in suits and ties, and singing together. There are at least 12 people in the group, ….The stage is set up with a microphone in the foreground, and a TV screen is visible in the background, likely displaying lyrics or information related to the performance. The audience is seated in the background, watching and enjoying the performance **…**
VCD: The image features a group of young men standing on a stage, dressed in suits and ties, and singing together. ... The stage is adorned with a large screen behind the group, which likely displays lyrics. The audience is seated in the background, appearing to watch and enjoy the performance. ..."

SID: The image depicts a group of young men performing on stage, wearing jackets and ties, and holding microphones. ... A TV screen is visible in the background, likely part of the stage setup. The audience is seated in the dark area facing the stage, observing the show. ..."
AVISC: The image features a group of young men standing on a stage, dressed in suits and ties, and singing together. …The stage is adorned with a screen, which could be used for displaying lyrics, visuals, or other information related to the performance. The audience is seated in the background, watching and enjoying the performance."The image captures a group of young men standing on a stage, dressed in semi-formal attire with ties, and singing as a choir. ... The background features a screen for visual effects. Additionally, the audience is seated in the venue, focused on the musical performance. ..."
DiVE: The image features a group of young men standing on a stage, dressed in suits and ties, and singing together. There are at least thirteen people in the group, with some standing closer to the front of the stage and others further back. They appear to be a part of a choir or a musical group, possibly performing at a concert or event.\n\nThe stage is adorned with a screen, which likely displays lyrics or other visuals related to their performance. The group appears to be well-prepared and focused on their performance, creating an engaging atmosphere for the audience.

Figure 5: **Qualitative comparison on POPE, MME, and CHAIR.** We compare the performance of DiVE with the other decoding methods (employed on **Qwen2.5-VL-7B** (Top) and **LLaVA-v1.5-7B** (Bottom)). Green and red text indicate correct predictions and hallucinations respectively.

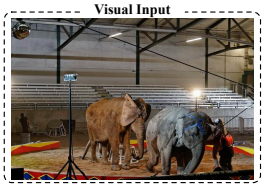
POPE

Visual Input

Question: Is there a cow in the image?
Label: Yes
Base (Qwen 2.5-VL-7B): "No, there is no cow in the image."
VCD: "There are sheeps in the image, but no cows."
SID: "No, there is no cow in the image."
AVISC: "No, there is no cow in the image."
DiVE: "There are sheep in the image, not cows."

MME-Hallucination

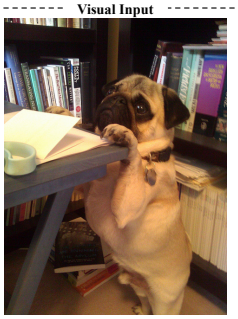
Visual Input



Question: Is the brown elephant in front of the gray elephant? Please answer yes or no.
Label: No
Base (Qwen 2.5-VL-7B): Yes.
VCD: Yes.
SID: Yes.
AVISC: Yes.
DiVE: Yes.

CHAIR

Visual Input



Question: Please describe this image in detail.
Ground truth words: [“dog”, “book”, “dining table”]
Base (Qwen 2.5-VL-7B): This image features a pug standing on its hind legs, leaning against a desk with its front paws resting on the edge. …The desk has some papers and a small blue bowl on it. **…** In the background, there is a bookshelf filled with books of various sizes and colors, …
VCD: The image features a pug standing on its hind legs, leaning its front paws against a desk. ... The desk appears cluttered with items, including loose papers and a small blue bowl, possibly containing water. The background is occupied by a large bookshelf filled with books of various colors, ..."

SID: This image displays a pug balancing on its hind legs, resting its weight on a desk. ... On the surface of the desk, there is a stack of papers and a distinct small blue bowl. ... Behind the dog, the room features a bookshelf that is densely packed with books, creating a studious atmosphere. ..."

AVISC: The image captures a pug standing upright on its hind legs, propped up against a desk. ... The desk surface holds several objects, such as a document and a small blue bowl near the edge. ... The background consists of shelves lined with books, suggesting the setting is a home office or library. ..."

DiVE: This image features a pug standing on its hind legs, leaning on a desk with its front paws. The pug appears to be looking up, possibly at the camera or someone off-frame. The desk is cluttered with various items, including a stack of papers, a small blue bowl, and a pen. … The background is filled with bookshelves, which are densely packed with books of various sizes and colors, …

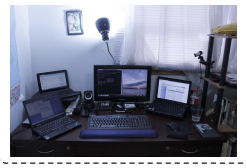
POPE

Visual Input

Question: Is there a traffic light in the image?
Label: No
Base (LLaVA-1.5-7B): Yes
VCD: Yes
SID: Yes
AVISC: Yes
DiVE: Yes

MME-Hallucination

Visual Input



Question: Is the light under the computer in the image? Please answer yes or no.
Label: No
Base (LLaVA-1.5-7B): Yes
VCD: Yes
SID: Yes
AVISC: Yes
DiVE: Yes

CHAIR

Visual Input

Question: Please describe this image in detail.
Ground truth words: [“person”, “traffic light”, “car”]
Base (LLaVA-1.5-7B): The image depicts a busy street with a red car driving down the road, surrounded by several other cars. There are multiple traffic lights positioned along the street, … and a couple of people are standing near the traffic lights.\n\nThe street is lined with buildings, giving it a town-like atmosphere. There are also a few benches placed along the sidewalk, providing seating for pedestrians. …
VCD: The image depicts a busy street scene with a red car driving down the road…A traffic light is visible on the left side of the street, and a stop sign is present on the right side. …There are also a few pedestrians walking along the sidewalk, adding to the bustling atmosphere of the scene.…The image depicts a busy street scene with a red car driving down the road... A traffic light is mounted on the left side, and a stop sign appears to be present on the right. ... Additionally, a few pedestrians are walking along the sidewalk, contributing to the busy atmosphere. ...
SID: The image illustrates a busy street with a prominent red car moving down the road, surrounded by traffic. ... Several traffic lights are positioned along the route, ... and a few benches are situated on the sidewalk, offering seating for passersby. ...
AVISC: The image portrays a busy street scene featuring a red car in motion... A traffic light is clearly visible on the left, while a stop sign is located on the right side of the intersection. ... There are also several pedestrians strolling along the sidewalk, adding life to the scene ..."

DiVE: The image depicts a busy street with a red car driving down the road, surrounded by other vehicles. There are several cars and a truck in the scene, with some cars positioned behind the red car and others in front of it. A traffic light is visible on the left side of the road, indicating that the traffic is being regulated.\n\nIn addition to the vehicles, there are multiple people walking along the sidewalk, some of them carrying handbags. …

Figure 6: **Error analysis on challenging samples.** We compare the performance of DiVE with the other decoding methods (employed on **Qwen2.5-VL-7B** (Top) and **LLaVA-v1.5-7B** (Bottom)). Green and red text indicate correct predictions and hallucinations respectively.