VES: VISUAL GROUNDING-AWARE STEERING FOR VISION LANGUAGE HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) achieve strong performance but are susceptible to hallucinating contextually plausible objects. Current LVLM steering methods mitigate hallucination by steering models away from these hallucination directions. However, these directions can encode semantic context shared by both the hallucinated objects and true image content. Consequently, we discover that the current methods suppress the true-content tokens after steering, a failure mode we term *Post-Steer Content-Suppression*. This problem arises because existing methods suppress hallucinatory directions without grounding their interventions in visual input. Building on this insight, we introduce a novel framework called *VES* (Visual-Evidenced Steering), where we actively steer the LVLMs towards directions supported by the image. First, *VES* identifies semantic regions within the image to construct a steering direction for each region, identifying visually grounded directions. It then weights these region directions according to their relevance to the current generated response and steer the model towards the aggregated direction. Across multiple LVLMs and four hallucination benchmarks, *VES* achieves state-of-the-art performance, showing that effective hallucination mitigation requires not only suppressing unsupported content but also reinforcing image-grounded evidence.

1 INTRODUCTION

Large Vision Language Models (LVLMs), such as LLaVA (Liu et al., 2023), InstructBLIP (Dai et al., 2023), Qwen-VL (Wang et al., 2024), and Molmo (Deitke et al., 2025; Clark et al., 2026), have achieved impressive performance across a wide spectrum of multimodal tasks such as image captioning, spatial reasoning. Despite this progress, these models remain vulnerable to visual hallucination, where they describe objects or attributes that do not appear in the image (Rohrbach et al., 2018; Li et al., 2023b). Such errors pose risks in high-stakes settings such as robotics and healthcare, making effective mitigation of hallucination essential to improve the trustworthiness of LVLMs.

There have been various works tackling visual hallucination mitigation during inference time. Some popular paradigms include contrasting output distributions during decoding (Leng et al., 2024; Huang et al., 2024), modifying the model's attention weights Huang et al. (2024); Liu et al. (2024c), or by steering hidden states in activation space (Liu et al., 2024b; Li et al., 2025; Yang et al., 2025a). LVLM's steering in particular is an exciting direction, requiring no further training or parameters updates, adds negligible inference cost. Prior works, for example, project features out of a hallucination subspace (Yang et al., 2025a), reinforce a global visual direction (Liu et al., 2024b; Li et al., 2025), or adapt the steering strength per token (Yin et al., 2026; Zhang et al., 2026).
Although effective, these methods primarily steer the model away from hallucination directions Li et al. (2025); Yin et al. (2026); Liu et al. (2024b); Dang et al. (2026), without considering whether that direction also supports genuine image content. Hallucinated objects are often contextually plausible and therefore semantically related to objects actually present in the image(Shoby et al., 2026). As illustrated in Fig. 1a, the direction associated with the hallucinated token *boat* can be close to that of the genuine token *surfboard*, because both are consistent with the same beach context. Steering away from *boat* can therefore unintentionally suppress *surfboard*. We term this failure mode *Post-Steer Content-Suppression*.

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

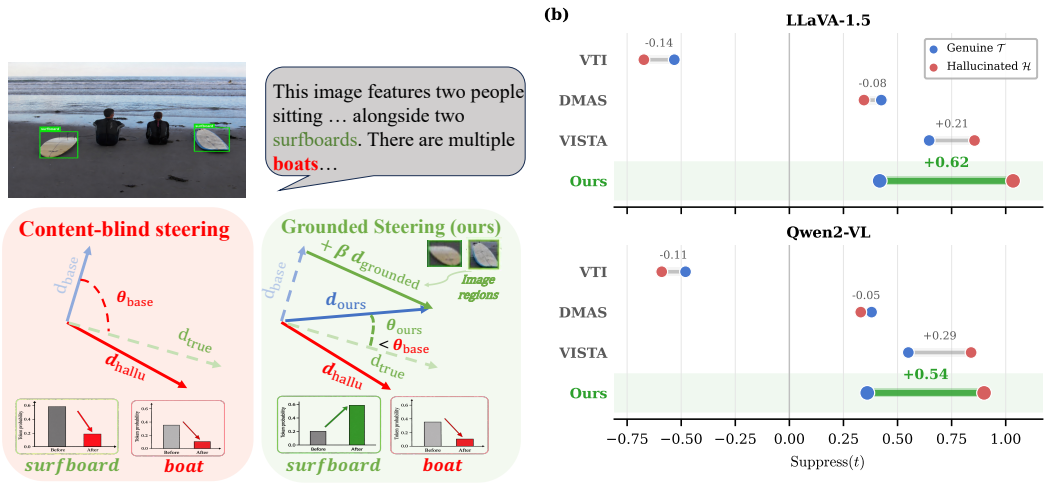


Figure 1: ***Post-Steer Content-Suppression* in existing steering methods. (a)** Illustrative comparison between our method and other LVLM steering works. Existing methods only focused on steering LVLMs away from a hallucinatory tokens' direction, which can correlate heavily with those of true tokens, thereby suppressing both. In contrast, our method integrate directions that support these true content, shifting generation toward image-supported content. **(b)** Suppression-score distributions for true tokens $\mathcal{G}$ (blue) and hallucinated tokens $\mathcal{H}$ (red), measured on 500 MSCOCO images. Existing methods distinguishes true and hallucinated tokens poorly in both LLaVa-1.5 and Qwen2-VL, leading to both of these tokens suppressed after steering.

Existing methods *significantly suppress both* hallucinated and true tokens with comparable magnitude. As shown in Fig. 1(b), current LVLM steering approaches suppress true tokens as much as hallucinated tokens. As such, they can suppress and miss true objects after steering, or let other hallucinatory tokens emerge instead. To correctly predict the true tokens, the logit of hallucinated tokens should be suppressed significantly lower than those of the true ones. However, current state-of-the-art methods suppress the true tokens after steering similarly from hallucinated tokens. (see Fig. 1b). This problem arises because they suppress hallucinatory directions *without grounding in image-supported contents*.

These observations motivate our approach *Visual-Evidenced Steering* (*VES*), a steering framework that utilize the image's regions as positive steering signals. At a high level, *VES* proceeds in two stages. First, we divide the image into semantic regions and construct a steering direction for each region. Second, during decoding, it weighs the regions directions based on how relevant they are to the generated context, then combines their directions to guide the model toward content supported by the image. We implement region discovery using an unsupervised segmentation algorithm, utilizing CLIP-based features clustering Shao et al. (2024), and dynamically estimate region relevance using a cosine similarity based computation.

Finally, we demonstrate that content-aware steering leads to state-of-the-art results across LLaVA-1.5, Qwen2-VL, and Molmo 2 Liu et al. (2024a); Wang et al. (2024); Clark et al. (2026) on CHAIR, POPE, MME, and AMBER (Rohrbach et al., 2018; Li et al., 2023b; Fu et al., 2026; Wang et al., 2023). Our contributions are as follows:

- To the best of our knowledge, we are the first to uncover and characterize the phenomenon we dubbed *Post-Steer Content-Suppression* stemming from steering methods neglect reinforcing the visual content, they suppress true tokens as strongly as hallucinated ones. We expose this behavior through a per-token analysis of post-steering logit changes across representative methods (Fig. 1).
- Building on the insight that corrective signals must come from the specific image region supporting each token, we propose *VES*, a training-free steering framework that discovers a per-image dictionary of region directions, and adaptively steers the model's generation process to be closer to these image-supported content.

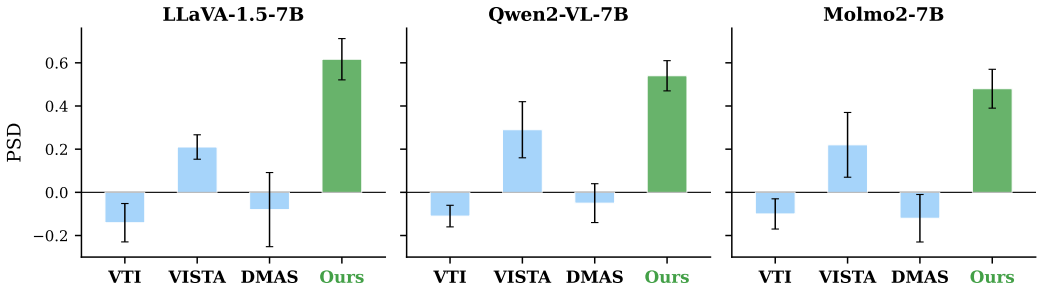


Figure 2: **Content-blind suppression in existing steering methods.** PSD on 500 MSCOCO images on LLaVA-1.5-7B, Qwen 2-VL-7B, Molmo 2 7B. We observe VTI, DMAS has small negative PSD values, demonstrating limited discrimination between true and hallucinated tokens. VISTA has a better but moderate PSD score, showing they have weaker discriminabilities for true and hallucinated tokens

- *VES* achieves state-of-the-art results on mutliple vision hallucination benchmarks across state-of-the-art LVLMs, demonstrating the effectiveness of our approach.

2 Content-Blind Steering Suppresses Genuine Tokens and Diminishes Hallucination Mitigation Capabilities

Finding 1: Content-blind steering suppresses true tokens alongside hallucinated tokens.

Existing steering methods commonly apply a direction designed to steer the models away from hallucinatory behaviors $\mathbf{v}^{(\ell)} \in \mathbb{R}^d$ at layer ℓ (Liu et al., 2024b; Li et al., 2025; Yang et al., 2025a; Yin et al., 2026). Whether learned globally or derived from an image-level representation, this direction is not selected from the regions relevant to the current token. Thus, steering away from hallucinated beach-related objects can also suppress a genuine object such as *surfboard* like in Figure 1a.

To measure this effect precisely, let $y_{1:T}$ be a model-generated description. Based on CHAIR's object ground truth labels (Rohrbach et al., 2018), we partition its object-token positions into a *true* set $\mathcal{G}$, whose referents are present in the image, and a *hallucinated* set $\mathcal{H}$, whose referents are absent. At position i, let $\ell_i^{\text{base}}(y)$ and $\ell_i^{\text{steer}}(y)$ denote the base and steered logits of candidate token y, respectively, conditioned on the same image and fixed prefix $y_{<i}$ under teacher forcing. We define the candidate-wise *suppression score* as

$$s_i(y) \;=\; \ell_i^{\text{base}}(y) \;-\; \ell_i^{\text{steer}}(y), \qquad s_i \;=\; s_i(y_i), \tag{1}$$

where $s_i > 0$ indicates suppression and $s_i < 0$ indicates promotion of the generated token. Selective steering should yield positive scores for hallucinated positions while leaving genuine positions unchanged or promoting them. We capture this behavior with *Post-steer Discriminability* (PSD), the difference between the mean scores of the two sets:

$$\text{PSD} \;=\; \bar{s}_{\mathcal{H}} - \bar{s}_{\mathcal{G}}, \qquad \bar{s}_{\mathcal{S}} \;=\; \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} s_i, \quad \mathcal{S} \in \{\mathcal{G}, \mathcal{H}\}. \tag{2}$$

PSD is measured in logits. A positive PSD means hallucinated tokens are suppressed more than genuine tokens; a near-zero PSD indicates little selectivity, and a negative PSD means true tokens are suppressed more.

Figure 2 shows near-zero discriminability for VTI and DMAS; their fixed directions shift genuine and hallucinated token logits by comparable amounts, and negative PSD indicates that genuine tokens are suppressed marginally more. VISTA improves PSD to $+0.21$ but remains below *VES* at $+0.62$. We refer to this failure as *Post-Steer Content-Suppression*.

Post-Steer Content-Suppression means that steering can substantially reduce the logits of true tokens alongside hallucinated ones. This phenomenon can prevent those true tokens from being predicted

by the model, allowing a plausible but hallucinated token to take its place.Consequently, this can decrease the performance on downstream hallucination mitigation. We therefore seek steering directions that suppress hallucinated tokens while preserving or increasing the logits of true objects that are supported by visual evidences.

Which representations efficiently encode true-object directions? To alleviate the true-token suppress problem, our goal is finding the representations strongly aligned with the true-object directions. We discovered that LVLM's object region tokens' representations itself encode strong visual-grounding directions. Steering with them surpasses using the last textual token's hidden states, as proposed in Liu et al. (2024b); Li et al. (2025); Yin et al. (2026). To the best of our knowledge, this paper is the first work to empirically identify that **object-region tokens** are the most optimal LVLM features to encode true objects' directions.

Finding 2: Object-region tokens efficiently encodes true-object directions

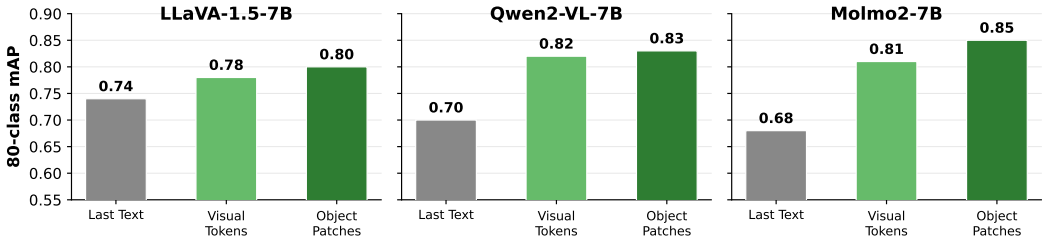


Figure 3: **80-class COCO object classification mAP at the best layer for each model (layer** 25**,** 22 **and** 35 **for LLaVA, Qwen 2 VL, and Molmo 2,respectively) , by token position.** Object visual patches achieve 0.85 mAP, outperforming the global visual-token mean (0.82) and last text tokens (0.74). Object identity is encoded most strongly at the spatial positions of the object.

In contrast to existing works where the visual capabilities of LVLM's hidden represenstations are studied layer-wise Wang et al. (2025); Li et al. (2025); Nguyen et al. (2026), we instead investigate the effect of tokens' positions on their visual information encoding capabilities. Our goal is to evaluate the steering alignment with the true-content directions of different tokens hidden representation. Specifically, we did 80-class linear probing on 3,000 MS COCO images for three different kinds of LVLM tokens' features: the last text tokens, the averaged hidden states of image tokens, and the averaged hidden states of image patches that belong to the predicted object's bounding box. The details of implementation are reported in the Appendix.

From Fig 3, we see that image tokens have significantly better visual capabilities than textual tokens, with image tokens achieving better mAP scores than textual tokens. Notably, an object's actual region is most useful for encoding that particular object, where we obtain 0.85 in mAP when probing the object's area hidden states. This finding motivates us to identify key regions within the image and use signals from those regions to ground the steering direction close to true content.

3 METHOD

Motivated by the findings in Section 2, we introduce *VES* (Visual-Evidenced Steering), a training-free method that grounds decode-time activation steering in the visual regions of each input image. Rather than applying a global direction shared across images and tokens, *VES* constructs an image-specific bank of region directions during prefill and dynamically combines the directions relevant to each decoding step. The full method is illustrated in Figure 4

Overview. Let an LVLM encode image I into P visual tokens. We denote the hidden state of visual token p at decoder layer ℓ by $\mathbf{h}_p^{(\ell)} \in \mathbb{R}^d$, and the hidden state of the token generated at step t by $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^d$. Our goal is to steer $\mathbf{h}_t^{(\ell)}$ away from hallucinatory directions while also being aligned with visual evidence that is present in I. *VES* proceeds in three stages: (i) **adaptive region discovery**, which groups visual tokens into semantic regions and constructs a layer-wise region bank; (ii) **dynamic region selection**, which estimates the regions relevant to the current decoding state; and (iii) **region steering**, which injects the selected visual evidence while preserving the hidden-state norm.

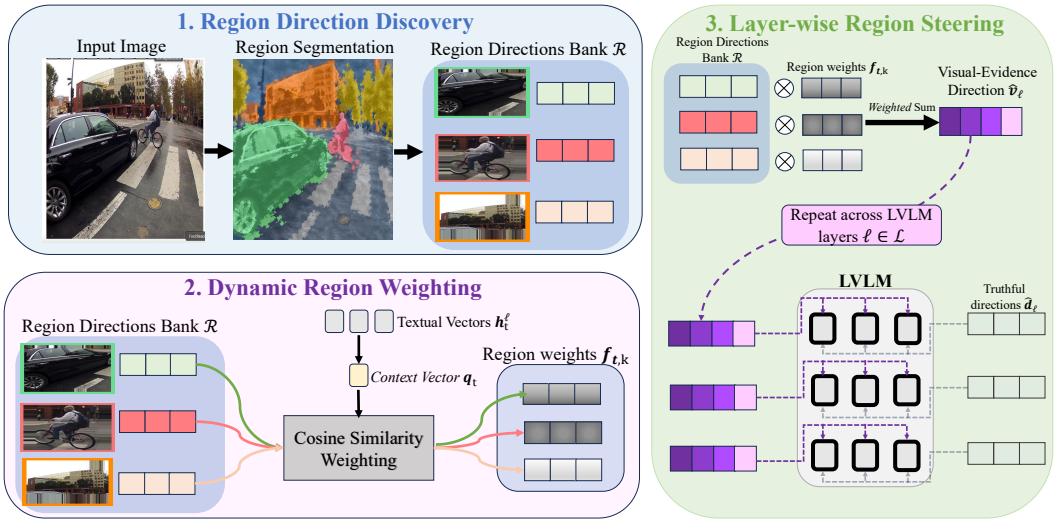


Figure 4: Our layer-wise region steering framework. **(1)** Region-specific directions bank $\mathcal{R}$ are extracted from semantic image regions obtained through unsupervised segmentation on CLIP features. **(2)** These region directions from $\mathcal{R}$ are dynamically weighted by pooling their similarities with the textual context q_t at current time step t to obtain the region weights $f_{t,k}$ at this time step. **(3)** The region directions $r_k^{(\ell)}$ at layer ℓ are aggregated with region weights $f_{t,k}$ at each selected LVLM layer (4 last layers in our paper) to steer generation toward relevant visual evidence.

3.1 Adaptive Region Discovery

As stated by *Finding 2* in Section 2, we first aim to identify the image's critical regions that encode it's visual content. During inference, we collect the visual-token states $\{\mathbf{z}_p^{(\ell_c)}\}_{p=1}^P$ from a late LVLM layer ℓ_c. We apply a CLIP-based unsupervised segmentation algorithm (Shao et al., 2024) to obtain a partition $R = \{R_k\}_{k=1}^K$, where $R_k \subseteq \{1, \ldots, P\}$ contains the patch indices assigned to region k. Since the CLIP features are obtained during the LVLM's forward pass, so minimal extra compute was added for this operation. The number of regions K is selected independently for each image, allowing the representation to adapt to scene complexity without category labels or task-specific training. For every region k and steering layer $\ell \in \mathcal{L}$, we average the corresponding visual states and normalize the result:

$$\mathbf{r}_k^{(\ell)} = \frac{\sum_{p \in R_k} \mathbf{h}_p^{(\ell)}}{\left\| \sum_{p \in R_k} \mathbf{h}_p^{(\ell)} \right\|_2}. \tag{3}$$

We refer to $\mathbf{r}_k^{(\ell)}$ as the direction of region k at layer ℓ. The collection $\mathcal{R} = \{\mathbf{r}_k^{(\ell)} : k = 1, \ldots, K, \ell \in \mathcal{L}\}$ forms an image-specific *region bank*. It is computed once during prefill and reused throughout decoding.

3.2 Dynamic Region Selection

The visual evidence required during generation changes after every step t with the evolving context. We therefore maintain an exponential moving-average query $\mathbf{q}_t$ at a designated layer ℓ:

$$\mathbf{q}_t = (1 - \alpha)\mathbf{q}_{t-1} + \alpha \mathbf{h}_t^{(\ell)}, \qquad \alpha \in (0, 1), \tag{4}$$

initialized with the final prefill state. At each new decoding step, we compare the normalized query $\widehat{\mathbf{q}}_t = \mathbf{q}_t / \|\mathbf{q}_t\|_2$ with every region direction at ℓ_s:

$$c_{t,k} = \left\langle \widehat{\mathbf{q}}_t, \mathbf{r}_k^{(\ell_s)} \right\rangle, \qquad f_{t,k} = \frac{\max(0, c_{t,k})}{\sum_{j=1}^K \max(0, c_{t,j})}. \tag{5}$$

The region weights $f_{t,k}$ requires an active region to have positive similarity to prevent weak matches from triggering an intervention.

3.3 Layer-wise Region Steering

VES combines two steering signals at each layer $\ell \in \mathcal{L}$: a precomputed direction from existing images that steer the model away from hallucinatory direction, inherited from Liu et al. (2024b), and a per-image region direction from the region bank.

Truthfulness direction. Following VTI Liu et al. (2024b), we precompute a fixed steering vector from $N = 50$ calibration samples. For each sample i, we record the last-token hidden state at layer ℓ with and without visual input and average the differences:

$$\hat{\mathbf{d}}^{(\ell)} = \frac{\sum_{i=1}^{N} \left(\mathbf{h}_{i,\text{img}}^{(\ell)} - \mathbf{h}_{i,\text{txt}}^{(\ell)}\right)}{\left\|\sum_{i=1}^{N} \left(\mathbf{h}_{i,\text{img}}^{(\ell)} - \mathbf{h}_{i,\text{txt}}^{(\ell)}\right)\right\|_2}. \tag{6}$$

This direction captures the general shift from text-only to visually-grounded generation and is computed once offline.

Per-image region direction. We aggregate k region directions identified from $\mathcal{R}$, weighted by their contextual relevance f_t from . equation 5 are shared across layers. At step t, we form the unit region direction:

$$\hat{\mathbf{v}}_t^{(\ell)} = \frac{\sum_{k=1}^{K} f_{t,k}\, \mathbf{r}_k^{(\ell)}}{\left\|\sum_{k=1}^{K} f_{t,k}\, \mathbf{r}_k^{(\ell)}\right\|_2}. \tag{7}$$

Content-aware Steering. We rotate the hidden state toward both directions while preserving its norm:

$$\mathbf{h}_{t,\text{VES}}^{(\ell)} = \left\|\mathbf{h}_t^{(\ell)}\right\|_2 \cdot \frac{\hat{\mathbf{h}}_t^{(\ell)} + \beta\, \hat{\mathbf{d}}^{(\ell)} + \lambda\, \hat{\mathbf{v}}_t^{(\ell)}}{\left\|\hat{\mathbf{h}}_t^{(\ell)} + \beta\, \hat{\mathbf{d}}^{(\ell)} + \lambda\, \hat{\mathbf{v}}_t^{(\ell)}\right\|_2}, \tag{8}$$

where $\hat{\mathbf{h}}_t^{(\ell)} = \mathbf{h}_t^{(\ell)} / \|\mathbf{h}_t^{(\ell)}\|_2$,$\lambda$ controls the strength of our proposed region steering, while β controls the global truthfulness signal from VTI. This double steering approach ensures the model doesn't diverge away from truthfulness directions while steering away from hallucinated ones.

4 Experiments

We evaluate *VES* on three LVLM architectures and four hallucination benchmarks. Following prior studies (Li et al., 2025; Wang et al., 2025), we first describe the evaluation protocol and baselines and then report the main quantitative comparisons.

4.1 Experimental Setup

Models. We conduct experiments on LLaVA-1.5-7B (Liu et al., 2024a), Qwen2-VL-7B (Wang et al., 2024), and Molmo 2-7B (Clark et al., 2026). These models cover distinct visual encoders and LLMs, allowing us to assess whether the proposed mechanism transfers across model archs.

Datasets and metrics. We evaluate open-ended object hallucination on CHAIR (Rohrbach et al., 2018), using 500 images from the MSCOCO validation set (Lin et al., 2014). We report sentence-level C_S, instance-level C_I, and object recall; lower CHAIR scores and higher recall are preferred. POPE (Li et al., 2023b) evaluates object-existence questions over its random, popular, and adversarial splits, each containing 3,000 questions, for which we report F1. For MME (Fu et al., 2026), we use the hallucination-related existence, count, position, and color subsets and report acciracu for each subset and their sum. We additionally use AMBER (Wang et al., 2023) for multidimensional hallucination evaluation.

Baselines. We compare against vanilla greedy decoding, nucleus sampling, and beam search, as well as representative training-free mitigation methods based on attention reallocation, contrastive or intermediate-layer decoding, and activation steering (Huang et al., 2024; Leng et al., 2024; Favero et al., 2024; Li et al., 2025; Wang et al., 2025; Liu et al., 2024b; Yin et al., 2026). This selection

Table 1: CHAIR evaluation on 500 MSCOCO images with a maximum generation length of 256 tokens. Lower C_S/C_I indicates fewer hallucinations, whereas higher Recall indicates better object coverage. Best and second-best results are bolded and underlined, respectively.

	LLaVA-1.5			Qwen2-VL			Molmo 2-7B		
Method	$C_S \downarrow$	$C_I \downarrow$	Recall↑	$C_S \downarrow$	$C_I \downarrow$	Recall↑	$C_S \downarrow$	$C_I \downarrow$	Recall↑
Vanilla	52.0	15.4	77.1	39.8	8.8	**68.6**	39.0	9.1	72.1
Nucleus	58.0	17.5	74.2	41.0	8.7	65.5	45.4	11.6	70.1
Beam	56.0	16.1	**77.8**	36.4	<u>6.9</u>	<u>67.8</u>	<u>34.6</u>	8.4	62.5
OPERA	53.2	15.0	75.9	39.6	8.3	66.7	40.4	8.8	72.1
VCD	58.0	17.1	<u>77.2</u>	40.8	9.8	67.6	50.6	12.0	72.3
M3ID	52.6	13.8	**77.8**	43.2	9.3	66.3	51.0	15.2	69.5
VISTA	46.4	13.2	74.4	<u>33.6</u>	7.2	67.7	42.0	9.7	71.8
DeCo	<u>41.2</u>	<u>11.3</u>	72.6	35.6	7.1	66.6	38.6	8.8	69.2
VTI	45.2	13.8	74.6	39.2	8.2	67.2	35.4	<u>7.6</u>	70.2
DMAS	42.0	11.4	73.7	34.3	7.6	67.3	35.0	7.9	<u>74.8</u>
VES	**38.8**	**9.2**	76.7	**32.6**	**6.6**	**68.6**	**33.8**	**7.1**	**75.6**

covers both output-space and representation-space interventions most closely related to our setting.

Implementation details. All methods are evaluated without additional training. For all hallucination mitigation methods, we use greedy decoding and the same prompt and generation settings across methods. We follow released implementations and their recommended hyperparameters for all baselines. The prompt templates we used for each tasks are included in the Appendix. CHAIR generations are capped at 256 new tokens. For our method *VES*, we set the region steering strength $\lambda = 0.12$, the truthfulness direction to be $\beta = 0.2$, the *Exponential Moving Average* query parameter $\alpha = 0.2$. We choose to steer the final 4 layers for each model classes. for the unsupervised region segmentation algorithm ClipTrase Shao et al. (2024), we used $\epsilon = 1.1$ and min_samples $= 7$, followed by adaptive merging of adjacent clusters with cosine similarity threshold $\tau_{\text{merge}} = 0.50$ and minimum region count $k_{\min} = 3$.

4.2 Experimental Results

CHAIR. Results in Table 1 show that *VES* consistently reduces hallucination across all three model families while preserving object coverage. We observe that *VES* outperforms state-of-the-art methods on the majority of metrics, with good hallucination reduction ($CHAIR_i$ score 9.2 compared with 11.3 from second-best method for LLaVA-1.5-7B) and effective object recall (68.6 for Qwen2-VL and 75.6 for Molmo 2-7B). It's worth noting that our method achieves the best balance between object recall and hallucination reduction, while certain methods were able to achieve reduced hallucination but sacrifices object recall such as DeCo (72.6) and VTI (74.6).

POPE. Table 2 demonstrates our model's performance on the object-existence benchmark. *VES* obtains the best F1 in seven of nine model-split settings and improves the average F1 over vanilla by 1.2 points on LLaVA-1.5, 2.2 on Qwen2-VL, and 0.7 on Molmo 2-7B. On the challenging adversarial split, F1 changes from 82.5 to 84.5, 85.8 to 86.5, and 86.6 to 87.8, respectively.

MME. Table 3 shows the comparative results on the multi-aspect benchmark MME. On LLaVA-1.5, *VES* raises the total from 631.7 to 661.6 and improves two of four subsets, including a 20.0-point gain on position. Qwen2-VL shows the same trend in Appendix 6, increasing from 708.3 to 721.0 and leading four of five columns. These broad gains suggest that the region bank reinforces localized objects, attributes, and spatial relations.

AMBER. On AMBER (Table 4), *VES* reduces CHAIR_S from 9.2 to 7.2, Hal from 61.5 to 52.6, and Cog from 6.8 to 4.5, while slightly raising Cover from 74.8 to 74.9. This demonstrates the effective balance of our method, achieving good hallucination mitigation results while also having large object coverage.

Table 2: POPE evaluation on the official random, popular, and adversarial splits (3,000 questions per split). We report F1 score (%); higher is better. Best and second-best results are bolded and underlined.

	LLaVA-1.5			Qwen2-VL			Molmo 2-7B		
Method	Rand↑	Pop↑	Adv↑	Rand↑	Pop↑	Adv↑	Rand↑	Pop↑	Adv↑
Vanilla	85.6	84.6	82.5	88.9	87.5	85.8	90.0	89.3	86.6
Nucleus	83.8	82.6	80.3	88.2	86.1	85.2	88.6	87.6	85.7
Beam	85.6	84.6	82.5	89.0	87.5	85.8	86.0	85.3	82.5
OPERA	85.4	83.1	81.0	87.4	85.2	85.7	89.9	89.1	86.3
VCD	85.3	83.3	81.7	87.3	85.8	84.8	90.4	88.8	86.4
M3ID	85.7	84.6	82.6	88.9	87.4	85.7	90.0	89.3	86.6
VISTA	**86.9**	85.4	83.0	91.3	89.0	**86.9**	90.0	89.1	86.2
DeCo	85.7	84.6	82.6	88.9	87.5	85.8	90.2	89.3	86.6
VTI	81.4	80.8	79.4	88.1	86.9	85.3	89.0	89.5	86.4
DMAS	85.6	82.4	82.8	87.4	85.2	84.8	90.1	89.2	86.3
VES	85.8	**86.1**	**84.5**	**91.8**	**90.5**	86.5	**90.6**	**89.7**	**87.8**

Table 3: MME results on LLaVA-1.5.

Method	Exist↑	Count↑	Pos↑	Color↑	Total↑
Vanilla	**195.0**	158.3	123.3	155.0	631.7
Nucleus	180.0	105.0	130.0	138.3	553.3
Beam	190.0	141.7	121.7	160.0	613.3
VCD	190.0	151.7	128.3	138.3	608.3
VISTA	185.0	**163.3**	118.3	150.0	616.7
DeCo	**195.0**	158.3	123.3	155.0	631.7
VTI	190.0	153.3	128.3	151.7	623.3
DMAS	**195.0**	156.7	128.3	160.0	640.0
VES	**195.0**	153.3	**143.3**	**170.0**	**661.6**

Table 4: AMBER results on LLaVA-1.5.

Method	CH_S ↓	Cover↑	Hal↓	Cog↓
Vanilla	9.2	74.8	61.5	6.8
Nucleus	10.4	74.1	67.9	7.6
Beam	9.0	**75.2**	60.8	6.5
VCD	9.7	73.6	60.2	5.9
VISTA	8.1	75.0	58.6	6.2
DeCo	7.9	74.4	57.9	5.6
VTI	7.6	73.8	56.8	5.3
DMAS	8.7	74.5	59.7	5.8
VES	**7.2**	74.9	**52.6**	**4.5**

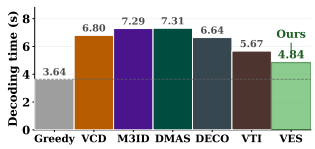


Figure 5: Latency of different mitigation methods

Efficiency Comparison. We compare the inference efficiency of different LVLM inference-time hallucination mitigation methods with our *VES* in Figure 5. We measure the latency for the image captioning task $CHAIR$, with a maximum of 256 tokens. Results are measured on a single Nvidia RTX A6000 GPU. We can observe that our method is generally efficient, only taking about 4.84 seconds/sample, which is only a 1.33 times overhead compared with greedy decoding, while outperforming other state-of-the-art methods both in performance and latency.

Qualitative analysis. Figure 6 illustrates an exemplar case where *VES* succeeded in its purpose. In this image, *VES* was able to boost the true token *"microwave"* significantly, while VTI decreased its logits. Thus, in the generated caption, our *VES* was able to mention this object while VTI missed it. In contrast, we were able to suppress the hallucinated token *"vase"*, which VTI hallucinated. In addition, we were able to mention unique objects that VTI didn't mention like *oven, chair, microwave*.

Ablation studies. We perform ablation studies on the effect of different components in our VES method on a 200 images subset of $CHAIR$. Specifically, from VTI as the base method, we incrementally add the global average visual features as additional steering vector instead of aggregrated region directions , replacing $\hat{\mathbf{v}}_t^{(\ell)}$ in Equation 8. Then we introduce region discovery with cosine similarities weighting by only the latest text token's hidden states, replacing $\mathbf{q}_t$ with $\mathbf{h}_t^{\ell}$ in Equation 4, and finally using the contextual moving average (EMA) for contextual weighting $\mathbf{q}_t$ to obtain the full *VES* method. The results are reported in Fig. 7. We observe that just by introducing the global visual vector to steering alone significantly improved upon VTI, going from 13.8 to 10.7 in $CHAIR_i$ and 45.2 to 38.0 in $CHAIR_s$. Pooling region features dynamically

instead of averaging all patches also significantly improved performance, going from 38.0 to 36.5 in $CHAIR_s$. *Exponential Moving Average* (EMA) queries resulted in incremental improvements, from 36.5 to 35.7 in $CHAIR_s$ while still being computationally efficient.



VTI The image is a … features a **refrigerator**, a **sink**, a **vase** and a dining **table**. The **refrigerator** is located on the left side of the kitchen, while the **sink** is situated in the middle…

VES The image showcases a large kitchen with a tile floor and a dining **table** with a tablecloth. The **kitchen** is well-equipped with a refrigerator, **microwave**, **oven**, and **sink** …with one **chair** on the left side…

Figure 6: **Qualitative case, captions shortened.** *VES* promotes grounded tokens *"microwave"* (green) while VTI suppresses it. In contrast, VES suppressed the hallucinated tokens *"vase"* while VTI slightly boosted it.

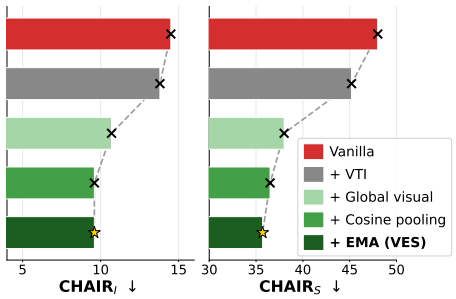


Figure 7: **Ablation of different components from *VES***. Visual Vector steering, Region Weighting, EMA Lower CHAIR_S is better.

5 Related Works

5.1 LVLM Hallucination

Visual hallucination refers to generated objects, attributes, or relations inconsistent with the visual evidence. CHAIR and POPE measure object hallucination via captions and existence polling (Rohrbach et al., 2018; Li et al., 2023b); MME offers broader perception evaluation, while AMBER and MMHal-Bench cover multiple hallucination types in discriminative and open-ended settings (Fu et al., 2026; Wang et al., 2023; Sun et al., 2024). Training-time methods improve faithfulness via factual feedback and preference alignment (Sun et al., 2024; Zhao et al., 2025b; Yang et al., 2025b; Liu et al., 2026; Yu et al., 2025). Training-free methods instead modify decoding with perturbed inputs Leng et al. (2024), visual grounding Favero et al. (2024), or intermediate-layer predictions (Leng et al., 2024; Favero et al., 2024; Chen et al., 2024; Wang et al., 2025; Wan et al., 2025), recalibrate attention or restore visual evidence internally (Huang et al., 2024; Liu et al., 2024c; Jiang et al., 2025; Zou et al., 2025; Nguyen et al., 2025), or add post-hoc verification from external vision models (Yin et al., 2024; Zhao et al., 2025a; Chen et al., 2024).

5.2 Activation Steering for Hallucination Mitigation

Activation steering edits internal representations without updating parameters. In language models, it intervenes on hidden states along behavior-related directions (Zou et al., 2023; Li et al., 2023a; Turner et al., 2025; Rimsky et al., 2024). LVLM methods learn hallucination or faithfulness directions, project out hallucination subspaces, or select sparse latent features (Liu et al., 2024b; Li et al., 2025; Yang et al., 2025a; Hua et al., 2025; Dang et al., 2026), and steering strength has also been adapted across decoding steps (Yin et al., 2026). However, none construct token-relevant directions from the grounded regions of each image. In contrast, we build an image-specific dictionary of region directions and dynamically inject the direction supported by the current visual context.

6 Conclusion

In this paper, we show that existing activation steering methods, which prioritizes steering away hallucinatory behaviors while neglecting visual grounding signals, can suppress genuine tokens as strongly as hallucinated ones, reducing object coverage without reliably eliminating unsupported content. Based on this finding, we introduce *VES*, which dynamically selects image-specific region directions to reinforce grounded tokens while suppressing unsupported alternatives. Extensive experiments across three LVLMs and four benchmarks demonstrate that *VES* consistently reduces hallucination while preserving descriptive coverage.

AI USE STATEMENT

We utilized generative AI tools to assist with manuscript drafting and editing, writing refinement, LaTeX formatting, figure preparation and visualization code. The authors take responsibility for the final content of this work, including all AI-assisted text, claims, code, and figures.

REPRODUCIBILITY STATEMENT

We have submitted the source code in the supplementary materials. The datasets used for the paper, the hyper-parameter settings can be found in Section 4.1. The Appendix A reports the prompt templates we used and the implementation details for motivational experiments in Section 2.

REFERENCES

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

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

Tiantian Dang, Chao Bi, Shufan Shen, Jinzhe Liu, Qingming Huang, and Shuhui Wang. Locate-then-sparsify: Attribution guided sparse strategy for visual hallucination mitigation. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026. URL https://openreview.net/forum?id=25n9KholuA.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104. IEEE, 2025.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Zhenglin Hua, Jinghan He, Zijun Yao, Tianxu Han, Haiyun Guo, Yuheng Jia, and Junfeng Fang. Steering LVLMs via sparse autoencoder for hallucination mitigation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10808–10828, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.572. URL https://aclanthology.org/2025.findings-emnlp.572/.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2024. doi: 10.1109/CVPR52733.2024.01274.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, 2025. doi: 10.1109/CVPR52734.2025.02328.

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

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=7BKcLeHQsm.

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

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

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

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*, 2024b.

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

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *Advances in Neural Information Processing Systems*, 38:111259–111284, 2026.

Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, Minh-Son To, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Localizing before answering: A benchmark for grounded medical visual question answering. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 7670–7678. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/853. URL https://doi.org/10.24963/ijcai.2025/853. Main Track.

Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Cam-Tu Nguyen, Minh Khoi Nguyen, Dang Huy Pham Nguyen, Anton van den Hengel, Johan Verjans, Phi Le Nguyen, et al. Beyond the global scores: Fine-grained token grounding as a robust detector of lvlm hallucinations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40235–40244, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

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

Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pp. 139–156. Springer, 2024.

Abin Shoby, Ta Duc Huy, Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Anton van den Hengel, Phi Le Nguyen, Johan W Verjans, and Vu Minh Hieu Phan. Overthinking causes hallucination: Tracing confounder propagation in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9194, 2026.

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. Aligning large multimodal models with factually augmented RLHF. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

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

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

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

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14635–14645. IEEE, 2025a.

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

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YtWZdwEG5K.

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

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19985–19995. IEEE, 2025.

Ruipeng Zhang, Zhihao Li, C.L.Philip Chen, and Tong Zhang. Steer where it matters: Token-level visual-sensitivity steering for LVLMs hallucination mitigation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=hCMNDdfzNb.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=w0xYx9CJhY.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond multimodal hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025b. doi: 10.1109/ICME59968.2025.11209377.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models, 2025. URL https://openreview.net/forum?id=tkg9XMFo0H.

A Full Implementation Details

A.1 Prompt Templates

Table 5 lists the exact prompt templates used for each benchmark.

Table 5: Prompt templates and maximum generation lengths for each benchmark.

Benchmark	Prompt Template	Max Tokens
CHAIR	`Please help me describe the image in detail.`	256
POPE	{`question`}\n`Answer the question using a single word or phrase.`	16
MME	{`question`} (includes "Please answer yes or no.")	16
AMBER	`Describe this image.`	512

A.2 Linear Probe Setup

To quantify how much object identity is encoded at different token positions, we train per-class linear probes on hidden states extracted from the final transformer layer.

Dataset. We use 3000 images and 500 images on MS COCO val2014 set to be used for training and evaluating the linear probes, separately. Ground-truth COCO bounding boxes of 80 different MS COCO annotated objects are mapped to visual token positions using each model's image preprocessing pipeline (e.g. shortest-edge resize and center-crop to a 24×24 patch grid for LLaVA-1.5).

Token positions. We extract representations from three token positions at the tokens' hidden states: (i) the *last text token*, (ii) the *average of all image tokens*, and (iii) *object patches*—the mean of visual tokens whose spatial positions fall within ground-truth bounding boxes.

Classifier. For each of the 80 COCO object categories we train a binary logistic regression probe with 200 epochs on ℓ_2-normalized hidden states. Classes with fewer than 1 test positive are excluded. We report the mean Average Precision (mAP) across all retained classes.

A.3 Post-steer Discriminability (PSD) experiment settings

True vs. hallucinated classification. Each generated caption is evaluated using the CHAIR metric (Rohrbach et al., 2018) to obtain per-word hallucination labels. Specifically, CHAIR identifies all noun phrases referring to COCO object categories (using a synonym dictionary that maps, e.g., "bike" $\rightarrow$ `bicycle`, "couch" $\rightarrow$ `sofa`). A mentioned object is labeled *true* if it appears in the ground-truth COCO annotation for that image, and *hallucinated* if it does not. For example, given an image annotated with {`dog`, `frisbee`}, the generated phrase "a dog catching a ball on the grass" yields `dog` as a true mention and `ball` as a hallucinated mention. Each labeled object word is then mapped back to its corresponding decode step via character-offset alignment, allowing us to associate every $\Delta\ell_t$ with a true or hallucinated label.

B Additional Experimental Results

B.1 Qwen2-VL Results on MME and AMBER

Table 6: MME results on Qwen2-VL. Higher is better.

Method	Exist↑	Count↑	Pos↑	Color↑	Total↑
Vanilla	195.0	160.0	**168.3**	185.0	708.3
Nucleus	190.0	150.0	148.3	168.3	656.6
Beam	195.0	160.0	163.3	185.0	703.3
VCD	190.0	156.7	156.7	178.3	681.7
VISTA	195.0	150.0	148.3	185.0	678.3
DeCo	195.0	160.0	163.3	185.0	703.3
VTI	193.3	158.3	160.0	181.7	693.3
DMAS	190.0	151.7	153.3	175.0	670.0
VES	**195.0**	**166.7**	166.0	**190.0**	**717.7**

MME. Table 6 shows the comparative results on the multi-aspect benchmark MME. On Qwen2-VL, *VES* raises the total from 708.3 to 717.7 and improves two of four subsets, including a 6.7-point gain on count.

Table 7: AMBER results on Qwen2-VL.

Method	$\text{CHAIR}_S \downarrow$	Cover↑	Hal↓	Cog↓
Vanilla	6.8	71.7	55.2	5.3
Nucleus	8.0	71.5	63.4	6.3
Beam	6.9	71.9	56.1	6.3
VCD	7.7	70.4	55.2	4.9
VISTA	6.7	71.8	57.5	6.4
DeCo	6.6	71.0	57.6	5.8
VTI	6.4	71.4	54.0	4.8
DMAS	6.6	71.3	54.7	5.0
VES	**5.9**	**72.5**	**50.6**	**3.7**

AMBER. On AMBER (Table 7), *VES* reduces CHAIR_S from 6.8 to 5.9, Hal from 55.2 to 50.6, and Cog from 5.3 to 3.7, while raising Cover from 71.7 to 72.5. This demonstrates an effective balance between hallucination mitigation and object coverage.

The GPT-4V-aided evaluation in Table **??** indicates that *VES* improves both the accuracy and detailedness of generated responses.

B.2 Additional Probing Experiments

We present the additional MS COCO layer-wise linear probing mAP results for the two models LLaVA-1.5-7B and Qwen 2-VL-7B in Figure 8. We observe that visual tokens have much better probing accuracy compared to the commonly used last text hidden states. We also see the object patches has significantly better visual information than global visual tokens across different layers. The visual capabilities of hidden states usually increases significantly around layer 20-25, which tracks with insights provided by previous works. Wang et al. (2025); Jiang et al. (2025).

B.3 Additional Ablation studies

We perform ablation studies on different values of the region steering strength λ for LLaVA-1.5-7B and steering layers for LLaVA-1.5-7B, Qwen 2-VL-7B and Molmo 2 7B on CHAIR benchmark. Figure 9 shows that greater steering strength significantly improves on hallucination mitigation,

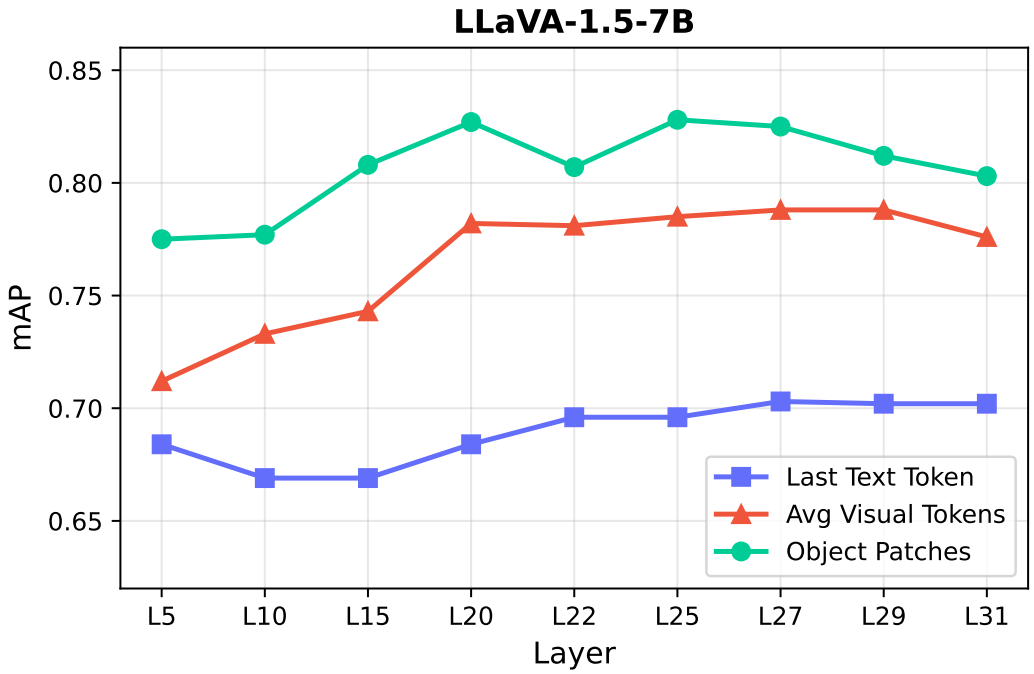


Figure 8: Layer-wise Probing mAP for LLaVA-1.5-7B

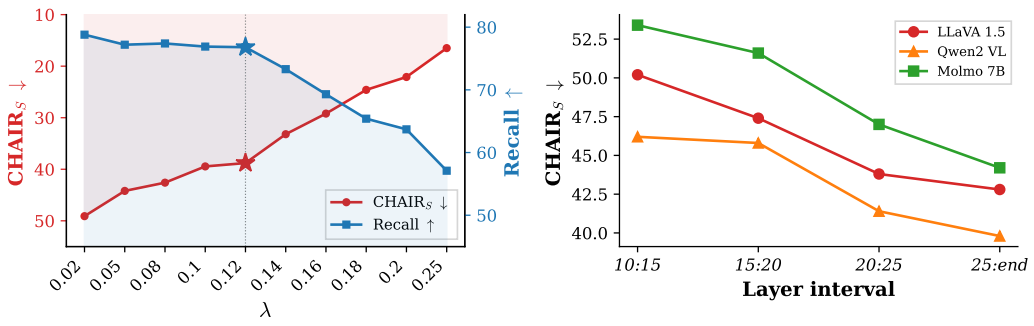


Figure 9: Ablation study on the steering coefficient λ and layer selection for the three model with $CHAIR_s$ and *Recall* reported on the *CHAIR benchmark*

however at the same time object recall, which reflects content diversity plummets. As steering strength grows larger, the model becomes more conservative and only describes details dominant in the image, making the generated content shorter and less diverse. In practice, we choose moderate values ($\lambda = 0.12$) to balance between hallucination mitigation and textual quality.
Steering later decoder layers also consistently improves performance, with the final layers interval strongest across all three model.

C Qualitative Results

C.1 Unsupervised segmentation results

We present the qualitative results of our unsupervised segmentation with CLIPTrase Shao et al. (2024) in Figure 10. We observe that while CLIPTrase doesn't return tight bounding boxes like supervised segmentation models, the algorithm still managed to capture the main semantic regions of interest quite well. For example, image (a) managed to separate the toilet with the trash can and

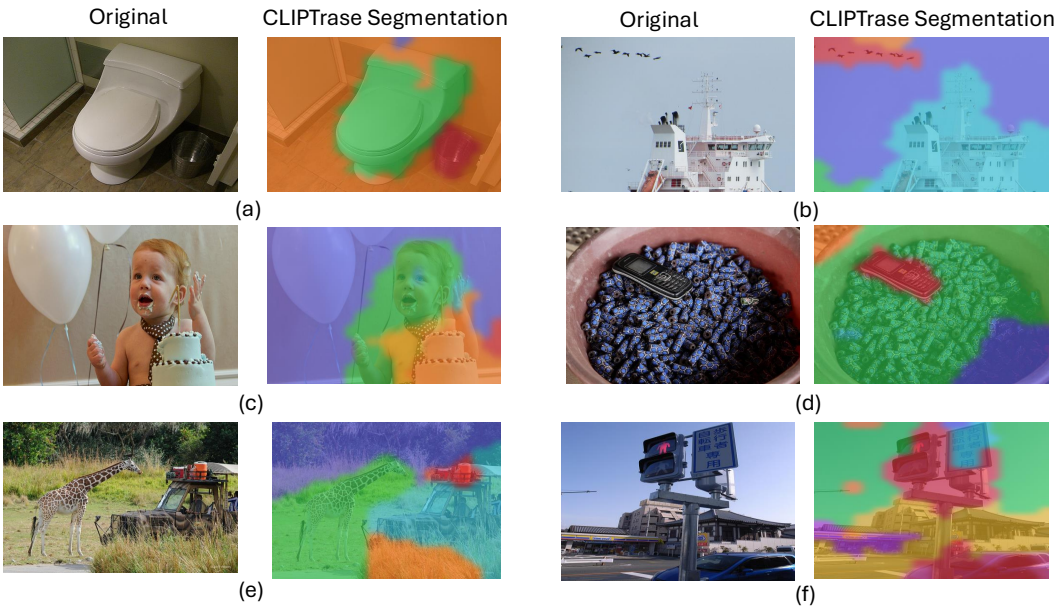


Figure 10: Qualitative Unsupervised segmentation with CLIPTrase.

the bathroom tiles, or image (f) where it was able to identify the region of the traffic lights, the sign beside the lights, the car passing by and background buildings.

C.2 VES Prediction Results

We visualize some predictions from *VES* for LLaVA-1.5-7B in Fig 11and Molmo 2 7B in Fig 12. We observe the effectivness of our method, improving upon hallucinated descriptions and occasionally help the model predict objects it missed when doing greedy decoding.



Greedy The image depicts a busy street scene with a large elephant walking down the road, carrying a man on its back. The elephant is positioned in the center of the scene, and the man is seated on a **bench** on its back. There are several people around the elephant, some of them riding bicycles. One **bicycle** is located near the left side of the scene, another is in the middle, and a third one is on the right side. A car can also be seen in the background, parked on the right side of the street. In addition to the people and the elephant, there are a few other objects in the scene, such as a **handbag** and a **backpack**, which are placed on the ground near the people.

VES The image depicts a busy street with a large elephant walking down the road, carrying a man on its back. The elephant is walking in front of a car, and several people are walking around, with some of them riding bicycles. There are several people in the scene, with some walking near the car and others walking on the sidewalk. A bicycle can be seen in the middle of the scene, and a car is parked on the right side of the street. The scene is bustling with activity, with people and vehicles sharing the road.",



Greedy The image features a dining table with a white plate filled with a colorful salad, including carrots and **broccoli**. The salad is placed on a white plate, which is positioned on the table. \n\nIn addition to the salad, there are several carrots scattered around the table, some of which are placed near the plate and others are spread out across the table. A fork is also visible on the table, likely used for enjoying the salad. \n\nThe table setting is complemented by a chair positioned on the left side of the table, and a person can be seen in the background, possibly preparing to enjoy the meal.

VES The image features a dining table with a white plate of food, including a salad with carrots and other vegetables. The plate is placed on a white plate, and the vegetables are arranged in a visually appealing manner. \n\nThere are several carrots on the plate, with some placed in the center and others on the sides. A fork is placed on the table, ready to be used to enjoy the meal. The dining table is surrounded by chairs, with one on the left side and another on the right side.\n\nIn the background, there is a person, possibly a person preparing the meal or someone who is about to enjoy the meal. A **bowl** is also visible on the table, possibly containing a side dish or a salad.

Figure 11: Qualitative Results for LLaVA-1.5-7B from Greedy Decoding and *VES*. We denote hallucinated objects in red and true objects Greedy Decoding in green.



Greedy The image captures a heartwarming birthday celebration for a young girl. She's seated at a table, blowing out candles on a white cake decorated with pink flowers. The girl is wearing a pink dress with red flowers and has long brown hair. To her left is an older girl with long blonde hair, wearing a blue shirt. On the right side of the table, there's another girl with long brown hair in a white dress with pink flowers. Behind them stands an adult woman in a black dress, likely the girl's mother. The table is set with a colorful tablecloth and various party items. There's a plate of pink candies, a **cup with straws**, and a stack of white napkins. Two plates featuring Disney princesses, Belle and Ariel, are also visible. The background shows a cozy living room setting with a lamp, a vase of flowers, and a window with brown blinds. The overall atmosphere is warm and festive, capturing a special moment in the girl's life.",

VES The image captures a heartwarming birthday celebration for a young girl. The scene is set in a cozy living room with a warm, inviting atmosphere. The focal point is a birthday cake on a table, surrounded by three girls. The birthday girl, a young girl with long brown hair, is seated at the table. She's wearing a pink dress and is blowing out the candles on the cake. The cake is decorated with pink flowers and has pink candles. To the left of the birthday girl is an older girl, likely her sister, with long blonde hair. She's wearing a blue shirt and is looking down at the birthday girl. On the right side is a younger girl with long brown hair, wearing a white dress with pink flowers. She's looking at the birthday girl with a smile. The table is set with a pink and white tablecloth, and there's a plate of pink candies in the foreground. A white chair is visible in the background, and a lamp with a white shade is lit, adding to the warm ambiance. The scene is filled with a sense of family and togetherness, capturing a special moment in the lives of these young girls. The colors in the image are predominantly pink, white, and brown, creating a soft and festive atmosphere



Greedy The image captures a professional tennis player on a blue hard court, likely at a major tournament such as the US Open. The player, who appears to be a man, is wearing a white shirt with red stripes and a white headband. He's holding a white towel with red stripes and the word \"OPEN\" printed on it, which he's using to wipe his face. In his right hand, he's gripping a tennis racket. The court is surrounded by a blue wall with advertisements for Emirates Airline and Panasonic. A large crowd of spectators can be seen in the background, filling the stands and creating an energetic atmosphere. To the right of the player, there's a cameraman in a yellow shirt and black pants, capturing the action on the court. The image is taken from a low angle, giving a sense of being on the same level as the player. In the foreground, there's **a white object that could be a ball** or a piece of equipment. The overall scene conveys the intensity and excitement of a professional tennis match, with the player appearing focused and ready for the next point.",

VES The image captures a professional tennis player on a blue hard court, likely at a major tournament such as the US Open. The player, who appears to be a man, is wearing a white shirt with red stripes and a white headband. He's holding a tennis racket in his right hand and a white towel with red stripes in his left hand, which he's using to wipe his face. The player is looking to his right, possibly at the scoreboard or another player. The court is surrounded by a blue wall with advertisements for Emirates Airline and Panasonic. In the background, there's a large crowd of spectators in the stands, many wearing white hats. A cameraman in a yellow shirt and black pants is visible, capturing the action on the court. The image is taken from a low angle, giving a sense of being on the court itself. There's a white object in the foreground, which could be part of the camera equipment or a piece of equipment on the court. The scene is vibrant and energetic, capturing the excitement of a professional tennis match in progress. The blue court, white lines, and the player's attire create a striking contrast against the blue wall and the crowd in the background.",

Figure 12: Qualitative Results for Molmo 2 7B from Greedy Decoding and *VES*. We denote hallucinated objects in red.