Modeling Visual Flow Leakage to Alleviate Hallucination in Large Vision-Language Models

Qi Bi[1], Jingjun Yi[2], Remco C. Veltkamp[1], and Albert Ali Salah[1]

[1] Utrecht University, Utrecht 3584CC, The Netherlands
{q.bi, R.C.Veltkamp, a.a.salah}@uu.nl
[2] University of Amsterdam, Amsterdam 1098XH, The Netherlands
{j.yi}@uva.nl

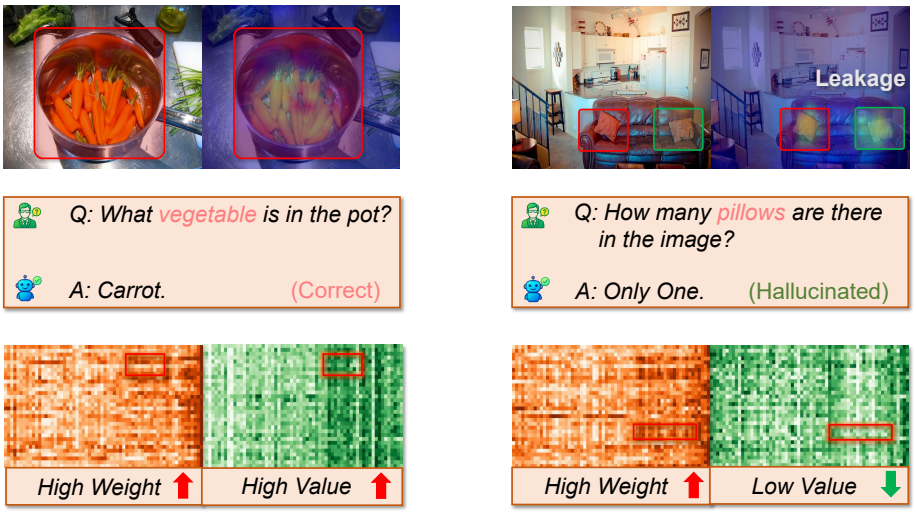
Abstract. Hallucination of large vision-language models (LVLMs) is a persistent challenge towards understanding and reasoning about an image. Prior remedies relied on contrastive decoding that compares multiple forward passes to improve the grounding. However, LVLMs can often look at the right image regions, and yet still produce text that is weakly grounded, which reveals a decoding-time failure mode that remains unaddressed. In this paper, we push this frontier forward through the new lens of Visual Flow Leakage (VFL), which quantifies the part of attention budget that is routed to visual tokens without contributing useful evidence for language completions. Based on this, we propose Flow Variance Reliability (FVR), a training-free decoder that preserves the generation efficiency while explicitly targeting VFL. Specifically, this decoder computes a per-head score from attention weights and projected value magnitudes to estimate visual value delivery. We then perform a reliability-guided intervention that mutes low-FVR heads in a layer and decode with the standard forward pass. Large-scale experiments across standard hallucination benchmarks show that the proposed method consistently reduces hallucination compared to existing training-free methods. We also hope our work could give insights to improve hallucination benchmarking with finer granularity.

Keywords: Large Vision-Language Models · Visual Flow Leakage · Hallucination Mitigation

1 Introduction

"I neither know nor think I know" – (in Plato, Apology 21d)

Large vision-language models (LVLMs) extend the reasoning ability of large language models (LLMs) to the visual domain [2,28,36], pushing the state of the art on diverse perception [4,5,52,56], understanding [25,43,48,53,58] and reasoning [21, 32, 42, 44, 45] tasks. Despite these advances, LVLMs frequently produce hallucinations, referring to general scenarios when the text it generates does not align with the content of the input image [34, 60]. Hallucination undermines the LVLM's reliability in real-world deployments [7, 14, 15, 41, 50, 59].



(a) Correct Sample (b) Hallucinated sample

Fig. 1: Attention weights (in orange matrix) are not sufficient to quantify the object contributions in an image. Such visual evidence can travel through different value paths (in green matrix) without injecting useful information when producing the logits. For all four matrices, the x- and y-axes denote token and layer dimensions, respectively. As shown in the token activation map [30], this phenomenon may result in a case that a LVLM 'sees' something, but yet is not 'aware' of the object and its contribution, causing a hallucination. We term this interesting phenomenon *Visual Flow Leakage*.

Recent works address hallucinations in LVLMs without additional training or multiple queries. A prominent paradigm is contrastive decoding, which compares outputs under perturbed inputs (*e.g.*, masked or noised images) to suppress language priors during inference [8, 11, 16, 23, 26, 27, 29]. While effective, such methods typically require two or more forward passes, amplifying both latency and memory costs [9,16,27]. More recently, stronger efficiency methods have been explored [49], better suited for real-time use. In either training-free paradigm, the hallucination mitigation hinges on how the perturbation is chosen and where the contrast is applied. As a result, these methods contrast the distributions but could not discern whether the attended visual information actually injects usable evidence, yielding failures where the model "sees" something relevant to the output, but loses this information somewhere in the process.

As shown in Fig. 1, the failure persists even when attention looks at relevant image tokens. The issue is not merely routing the probability mass, but *how much usable visual evidence travels from the attention weight (in orange matrix) to the value (in green matrix) that produces the logits.* As a result, there may be cases that lead to information leakage and hallucination, despite seemingly good attention deployment [1, 24]. We term this pathology **Visual Flow Leakage (VFL)**: attention is routed to visual tokens, but the routed value paths provide

weak or inconsistent evidence to the residual stream. This framing prioritizes *flow* over raw weights, and prompts us to formulate our research question: *Can we measure and correct the visual evidence flow, while preventing its leakage and incorrect language-prior completions?*

To push this frontier, we propose Flow Variance Reliability (FVR), a per-head score that favors heads whose visual value paths are aligned with the attention weights. Unlike conventional entropy-ratio head selection, this criterion is value-aware, marking heads whose visual values are dominated by brittle spikes as having higher risk of visual flow leakage. We subsequently apply a reliability-guided intervention. Heads with low FVR are muted in a layer, and the decoding proceeds with lightweight safety. A training-free method is accordingly proposed, which retains the speed of the fast decoders, while directly addressing VFL.

Notably, our work introduces the novel *Visual Flow Leakage* mechanism, which explains why "seen" areas can coexist with weak grounding. The proposed FVR can increase the weights of heads whose visual contributions are more coherent, but less variance-dominated. In contrast to entropy heuristics or multi-query contrastive families [10, 11, 20, 57], FVR is value-aware and quantifies the attention flows better than raw weights. We argue this design is more rational, as the attention output depends on not only the weights, but also the transformed values. Besides, it also aligns with the recent evidence that some highly attended visual tokens contribute negligibly to classification [1, 24]. This combination remains training-free with low decoding overhead while directly targeting VFL.

Large-scale experiments are conducted on three representative LVLMs, namely, LLaVA-1.5 [35], InstructBLIP [12], and Qwen-VL [2], and across multiple hallucination benchmarks, including POPE [31], CHAIR [42], and MME-Hallucination [17]. FVR maintains efficiency comparable to ONLY (3.42 s versus 3.70 s per image) and achieves a lower CHAIR_S of 49.6 versus 49.8. Our work also gives insights to improve hallucination benchmarking by more granularities, *e.g.*, finer-grained hallucinations and diversified distributions.

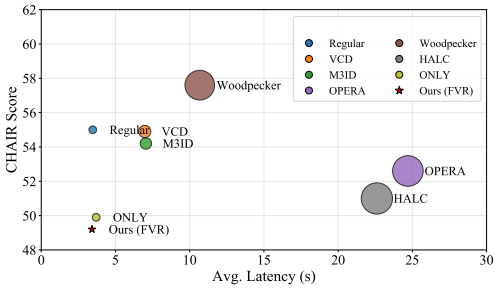


Fig. 2: Performance vs. efficiency comparison between the proposed method and existing methods. Y-axis: CHAIR score (lower is better; plotted descending); X-axis: Avg. Latency (s); Bubble size corresponds to GPU memory. All data are collected on a single RTX A6000 Ada GPU for fair evaluation [49]. Numerical computation outcomes are reported in Supplementary Material.

Our contributions can be summarized as follows.

– We propose *Visual Flow Leakage*, a novel value-path and flow view of hallucination, which explains why raw attention can "see" without grounding.

- We propose Flow Variance Reliability (FVR), a simple yet effective reliability score that measures gain-weighted visual value delivery for reliable head selection.
- We propose a reliability-guided intervention method to alleviate hallucinations in LVLMs, outperforming existing training-free methods while maintaining efficiency.

2 Related Work

2.1 LVLM and Hallucinations

Modern LVLMs [2,12,28,35] significantly boost the performance of various downstream computer vision tasks [4, 5, 25, 48, 52, 53, 58]. However, LVLMs still encounter hallucination, referring to the cases when the generated text does not align with the content in the image [34]. Their inherently dominant language priors can override image evidence under distribution shift, yielding object- and attribute-level hallucinations [8, 10, 11, 29].

2.2 Hallucination Mitigation

Prior works address the hallucination problems by instruction-tuning on curated data, adapting alignment strategies, and via preference optimization [12, 16, 35, 41, 59], but require substantial computing power and curation. Variants include additional grounding losses [47,51] or direct preference optimization [40, 55], which are costly and model-specific. To alleviate the computational cost, training-free, one-query route methods have drawn increasing attention. One research line follows the contrastive decoding paradigm, which intervenes at inference level, without weight updates [8,16,19,20,23,57]. Newer variants contrast instructions or language priors [14,15,37,50,51,59]. Still, these approaches often require multiple model passes and increase latency. Recently, a one-layer intervention during decoding has been proposed to address these issues efficiently [49]. Our work follows this line, but different from existing approaches, proposes a flow-aware reliability measure that directly scores visual contributions along the value paths.

2.3 Attention & Information Transport

Attention weights in a Transformer alone are insufficient to quantify contributions to the target task. Prior analysis formalizes the attention transport across layers via vector norms along value paths [1], gradient-style signals [6] and head compositions [38]. Among the many aspects of LLMs and LVLMs, one research line studies the observation of *attention sinks*, where some highly attended tokens contribute little to the final prediction but link to massive activation in specific hidden-state dimensions [13, 18, 22, 54]. Other related works associate layer-localization with attention budget misallocation [46]. The proposed visual

flow leakage (VFL) concept is fundamentally different from an attention sink. It is defined by mismatches between the attention route probability mass and the corresponding *value-to-logit flow*, which is dimension-agnostic and value-aware. VFL can pinpoint to cases with discrepancies in attention and logit contributions.

3 Preliminaries: Visual Flow Leakage

3.1 Notation & Definition

To reduce hallucinations in LVLMs, we focus on the case where the attention deployment is reasonable but information does not reach next-token logits. This failure mode arises because—even under high attention weights—the routed signal traverses heterogeneous *value paths*, and only a fraction of that routed mass converts into an effective update at the output. To quantify this *Visual Flow Leakage* (VFL) phenomenon, we adopt a mechanistic lens coupled with a computable reliability indicator.

Let i, j, h and ℓ denote the indices of the text position, visual position, head and layer of a LVLM (*e.g.*, LLaVA-1.5 [35]). Let $V^{(\ell,h)} \in \mathbb{R}^{T\times d_v}$ and $W_{(\ell,h)}^{OV} \in \mathbb{R}^{d_v\times d_o}$ denote the pre-projection values and per-head value-to-output projection, where T is the context length and $d_k/d_v/d_o$ is the key/value/output head dimension. Also let $q_i^{(\ell,h)} \in \mathbb{R}^{d_k}$ and $K^{(\ell,h)} \in \mathbb{R}^{T\times d_k}$ form the attention logits. Define $v_j^{(\ell,h)} = V_j^{(\ell,h)} W_{(\ell,h)}^{OV} \in \mathbb{R}^{d_o}$. The attention weight $w_{i,\cdot}^{(\ell,h)} \in \mathbb{R}^T$ is

$$w_{i,\cdot}^{(\ell,h)} = \mathrm{softmax}\Big(\frac{q_i^{(\ell,h)} K^{(\ell,h)\top}}{\sqrt{d_k}}\Big), \tag{1}$$

and the head output is $\alpha_i^{(\ell,h)} = \sum_j w_{i,j}^{(\ell,h)}\, v_j^{(\ell,h)} \in \mathbb{R}^{d_o}$.

We say Visual Flow Leakage (VFL) occurs at (ℓ, h, i) when the routed attention mass does not translate into an effective value update in the stream. Two ingredients govern this translation:

- the routing weights $w_{i,j}^{(\ell,h)}$, which decide where the probability mass flows;
- the path gain $G_j^{(\ell,h)} \triangleq \|v_j^{(\ell,h)}\|_2$, which decides how much visual information actually propagates into the stream.

3.2 Norm-aware Reliability Indicator

Our objective is to quantify VFL within the forward pass. The value in $v_j^{(\ell,h)}$ can be composed as $v_j^{(\ell,h)} = G_j^{(\ell,h)}\, \hat{v}_j^{(\ell,h)}$, where $G_j^{(\ell,h)} = \|v_j^{(\ell,h)}\|_2$ and $\|\hat{v}_j^{(\ell,h)}\|_2 = 1$. Then, the computation of $y^{(\ell,h)}$ is re-formulated as $\alpha_i^{(\ell,h)} = \sum_j w_{i,j}^{(\ell,h)} G_j^{(\ell,h)}\, \hat{v}_j^{(\ell,h)}$.

We measure the visual flow reliability by an indicator $N_i^{(\ell,h)}$, which aggregates gain-weighted routing mass as a norm, defined as

$$N_i^{(\ell,h)} = \sqrt{\sum_j \big(w_{i,j}^{(\ell,h)}\, G_j^{(\ell,h)}\big)^2}, \tag{2}$$

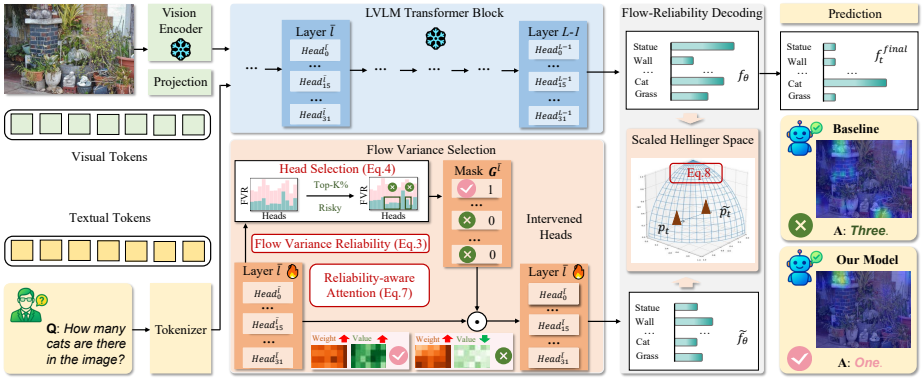


Fig. 3: Overview of the proposed Flow Variance Reliability (FVR) framework. Based on the FVR score (in Sec. 4.1) that models the visual flow leakage, it consists of two key components, namely, Flow Variance Selection (FVS, in Sec. 4.2) and Flow-Reliability Decoding (FRD, in Sec. 4.3), enabling training-free and computationally efficient LVLM hallucination mitigation.

where a high $N_i^{(\ell,h)}$ indicates a strong and high-energy routed signal for that head at position i. In contrast, a low $N_i^{(\ell,h)}$ indicates a weak and uncertain routed value, flagging instability and susceptibility condition for VFL. This indicator N will serve as our value-aware surrogate to identify VFL-prone heads and guide decoding in later sections.

4 Method

4.1 Flow Variance Reliability

Existing training-free hallucination mitigation methods can contrast the original distribution and the perturbed distribution well, which already allows a LVLM to focus on the key visual evidence (*e.g.*, object, landmark). However, how much trustworthy visual evidence reaches the stream when decoding remains to be an open question. Our observation of visual flow leakage (VFL) also raises this issue. The Norm-aware Reliability Indicator N (in Sec. 3.2) provides a feasible path to quantify the reliability of the attention in the transformer layers, but how to standardize it to select the reliable heads remains to be handled.

Based on the above discussion, the head selection in the proposed method relies on the Norm-aware Reliability Indicator N (Eq. 2). We say a head is *reliable* / *unreliable* when exhibiting high / low N value. We therefore define Flow Variance Reliability (FVR) $s_{\text{FVR}}^{(\ell,h)}$, a per-head reliability score, given by

$$s_{\text{FVR}}^{(\ell,h)} = \mathbb{E}_{\,i\in\mathcal{I}_t}\left[N_i^{(\ell,h)}\right], \tag{3}$$

where $\mathcal{I}_t = \{t\}$ denotes the current decoding position in our default setting.

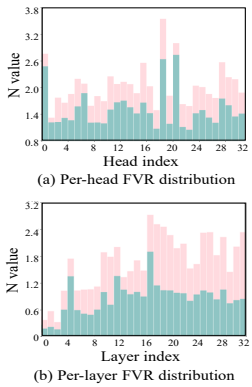


Fig. 4: Distribution of (a) Per-head FVR score and (b) Per-layer FVR score. Teal and magenta denote the hallucinated sample and the correct sample, respectively.

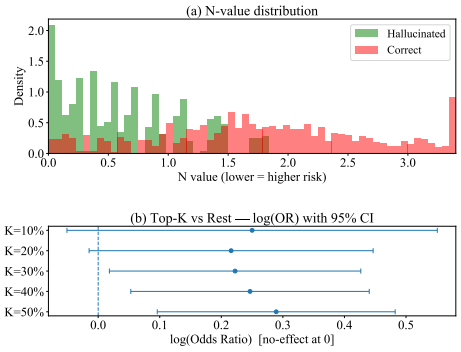


Fig. 5: (a) N-value distribution of hallucinated (red) vs. correct (pink) samples in POPE. (b) Forest plot when K is set to different ratios (95% CI). If the dot (blue) is to the right of 0, OR> 1, it indicates a positive effect.

A higher $s_{\text{FVR}}^{(\ell,h)}$ indicates robust value delivery while a lower $s_{\text{FVR}}^{(\ell,h)}$ indicates higher leakage risk. We therefore consider keeping the heads with higher $s_{\text{FVR}}^{(\ell,h)}$ while muting the heads with lower $s_{\text{FVR}}^{(\ell,h)}$ to reduce hallucinations. Fig. 4 visualizes the FVR score distributions, while Fig. 5 reports the corresponding N-value statistics and intervention effect.

4.2 Flow Variance Selection

Recent works have made significant progress on the selection of attention heads to address hallucination problems. Based on the distribution distortion, it becomes possible to select the heads [49]. While efficient, how much trustworthy visual evidence reaches the stream when decoding remains unaddressed. Therefore, our work aims to maintain the efficiency while further selecting the attention heads that substantially contribute to the subsequent stream and the decoding.

Let $\mathcal{H} \subseteq \{1, \ldots, H\}$ denote the head index within a single layer $\tilde{\ell}$. We use $s_{\text{FVR}}^{(\tilde{\ell},h)}$ to identify and mute the bottom-K% heads with the weakest value-path delivery. In other words, these heads are divided into two groups. The first group, G_{risky}, comprises the heads from the lowest K% FVR score distribution. The second group, $G_{reliable}$, comprises the heads from the highest (1-K)% FVR score distribution. By default we set K to be 20. This selection forms a masking matrix $\mathbf{G}^{(\tilde{\ell})}$, defined as

$$\mathbf{G}^{(\tilde{\ell})} = \text{diag}\big(m_1^{(\tilde{\ell})}\mathbf{I}_d, \ldots, m_H^{(\tilde{\ell})}\mathbf{I}_d\big) \in \{0,1\}^{Hd \times Hd}, \tag{4}$$

where $\mathbf{I}_d$ denotes the identity matrix, and the per-head mask $m_h^{(\tilde{\ell})}$ is defined as

$$m_h^{(\tilde{\ell})} = \begin{cases} 1, & h \in G_{\text{reliable}}, \\ 0, & h \in G_{\text{risky}}. \end{cases} \tag{5}$$

Let H_t^ℓ be the layer input and let $\tilde{a}^{(\tilde{\ell},h)}$ denote the output of head h before masking. The FVR-edited output is

$$\tilde{H}_t^\ell = \big[\tilde{a}^{(\tilde{\ell},1)}, \ldots, \tilde{a}^{(\tilde{\ell},H)}\big]\, \mathbf{G}^{(\tilde{\ell})} W_{\tilde{\ell}}^O, \tag{6}$$

where low-FVR heads are masked before the edited representation is propagated through the remaining layers. Here, $[\cdot,\cdot]$ denotes the concatenation operation. We observe that after muting and channel concatenation, the embedding remains dense, and the muting does not add many zeros. Besides, there is a residual connection in the Transformer layer followed by a layer norm and MLP. This compensation prevents the feature to collapse into zeros and reduces the redundancy in attention heads, ensuring numerical stability in practice (also see [39] for additional justification).

4.3 Flow-Reliability Decoding

Even when an LVLM looks at the right visual tokens, the routed attention mass may fail to materialize as aligned value contributions in the stream, creating Visual Flow Leakage (VFL). Our key idea is to make decoding flow-aware: we modulate the contribution of the edited logits by a reliability signal derived from FVR. We use the Hellinger discrepancy [3,33] as a bounded and stable measure of disagreement between the base and edited token distributions, allowing FRD to back off when the edit is unreliable.

At the t-th decoding step, let $\mathcal{S}$ denote the vocabulary, and let $f_\theta, \tilde{f}_\theta \in \mathbb{R}^{|\mathcal{S}|}$ denote the base and FVR-edited logits, respectively (Sec. 4.2). The corresponding distributions can be computed as $p_t = \text{Softmax}(f_\theta)$ and $\tilde{p}_t = \text{Softmax}(\tilde{f}_\theta)$, respectively.Then, we devise a scaled Hellinger discrepancy to compare next-token probability distributions on FVR, given by

$$\Delta_H(t) \;=\; 2\left(1 - \sum_{y\in\mathcal{S}} \sqrt{p_t(y)\,\tilde{p}_t(y)}\right), \tag{7}$$

and we define $\bar{\Delta}_H$ to normalize the gap to $[0,1]$ so it can act as a soft confidence estimator, computed as $\bar{\Delta}_H(t) \;=\; \frac{1}{2}\,\Delta_H(t)$. Then, let $\{s_{\text{FVR}}^{(\tilde{\ell},h)}\}_{h=1}^H$ be per-head reliability on the chosen layer $\tilde{\ell}$. We compress them into a single reliability scalar, given by

$$r_t \;=\; \frac{s_{\text{FVR}}^{(\tilde{\ell},h)}}{\sum_{h=1}^H s_{\text{FVR}}^{(\tilde{\ell},h)} + \varepsilon} \;\in (0,1), \tag{8}$$

where ε is a numerical stability hyper-parameter and by default set to be 10^{-12}. A higher r_t indicates a more reliable flow that has lower risk of visual flow leakage, and vice versa.

Finally, to devise a threshold-free blending, we form a convex weight that increases with reliability and decreases with discrepancy, given by

$$\begin{aligned} w_t &= \mathrm{clip}\big(r_t\,(1-\bar{\Delta}_H(t)),\,0,\,1\big), \\ f_t^{\mathrm{final}} &= (1-w_t)\,f_\theta \;+\; w_t\,\tilde{f}_\theta, \end{aligned} \tag{9}$$

where $\mathrm{clip}\big(r_t\,(1-\bar{\Delta}_H(t)),\,0,\,1\big) = \min\{\max\{r_t\,(1-\bar{\Delta}_H(t)), 0\}, 1\}$ clamps to $[0,1]$, and $y_t \sim \mathrm{Softmax}\big(f_t^{\mathrm{final}}\big)$. In this implementation, the edits are trusted when visual flow is reliable and consistent with the base (r_t high, $\bar{\Delta}_H$ low).

4.4 Training-Free Design & Framework Overview

Fig. 3 gives an overview of the proposed framework. Given a frozen LVLM, we insert a single reliability-guided edit at layer $\tilde{\ell}$: compute $s_{\mathrm{FVR}}^{(\tilde{\ell},h)}$ on-the-fly, mute low-reliability heads via Eq. 4, and propagate the edited multi-head output through the unchanged stack to obtain $\tilde{f}_\theta$ (Eq. 6). FVR favors heads with stronger gain-weighted value delivery, directly targeting Visual Flow Leakage. The blend in Eq. 9 then dynamically trusts the edited logits only when the measured flow is reliable and distributionally consistent, yielding training-free hallucination mitigation with computational efficiency. Specifically, it requires only quantities in an attention layer, namely per-head value norms $G_j^{(\ell,h)}$, unit directions $\hat{v}_j^{(\ell,h)}$, and attention weights $w_{i,j}^{(\ell,h)}$.

5 Experiments

5.1 Datasets & Baselines

Datasets. POPE [31] is a yes/no question benchmark to quantify object-level hallucinations. It consists of 27,000 query-answer pairs, and involves three sampling strategies, namely, random, popular, and adversarial. **CHAIR** [42] is an open-ended image captioning benchmark. It comprises 500 MS COCO validation images, focusing on object hallucination rates in generated captions. **MME-Hallucination** [17] is a comprehensive suite with four subsets evaluating object existence, object count, position, and color hallucinations.

Baselines. We compare the proposed FVR (denoted as Ours) with the state-of-the-art training-free hallucination mitigation methods, including VCD [27], M3ID [16], Woodpecker [57], HALC [10], DoLa [11], OPERA [19], ONLY [49], INTER [14], ICT [8], and fuzzycd [23]. Unless otherwise specified, the compared results are taken from [49]. For fair evaluation, we follow the default configuration and impose the proposed FVR on the first layer [49] unless otherwise specified. Besides, the responses are generated using stochastic sampling unless otherwise specified. INTER [14], fuzzycd [23] and ICT [8] are re-implemented by us with

Table 1: Results on the POPE [31] benchmark (object presence queries). Evaluation metrics are Accuracy and F1-score. Higher (↑) is better. Top three results are highlighted as **best**, second and third, respectively.

		MS-COCO						A-OKVQA						GQA					
		LLaVA-1.5		InstructBLIP		Qwen-VL		LLaVA-1.5		InstructBLIP		Qwen-VL		LLaVA-1.5		InstructBLIP		Qwen-VL	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Random	Regular	83.13	83.44	83.07	83.08	85.23	83.09	81.90	83.53	80.63	81.92	86.40	85.07	82.23	84.03	79.67	80.99	85.10	83.87
	VCD [27]	87.00	87.15	86.23	85.88	87.03	85.45	83.83	85.34	84.20	85.00	87.93	86.96	83.23	85.05	82.83	83.56	87.00	86.16
	M3ID [16]	87.50	87.52	86.67	86.41	86.40	84.50	84.67	85.97	85.43	86.23	87.50	86.32	84.20	85.77	83.07	83.87	87.07	86.16
	ONLY [49]	89.70	89.10	**89.23**	**88.89**	88.90	87.85	**86.07**	**87.14**	**88.57**	**88.94**	89.47	89.22	**86.70**	**87.83**	86.17	86.63	88.03	87.80
	Ours (FVR)	**89.87**	**89.90**	87.10	87.24	**89.33**	**88.57**	85.77	86.99	85.50	86.10	**89.90**	**89.70**	86.10	87.47	**86.23**	**87.00**	**88.87**	**89.18**
Popular	Regular	81.17	82.08	77.00	78.44	84.53	82.58	75.07	78.77	75.17	77.91	85.77	84.49	73.47	77.84	73.33	76.26	80.87	80.33
	VCD [27]	83.10	83.94	80.07	80.89	85.87	84.27	76.63	80.19	78.63	80.72	87.33	86.34	72.37	77.58	76.13	78.68	82.53	82.27
	M3ID [16]	84.30	84.95	80.97	81.85	86.07	84.30	77.80	80.91	78.80	81.00	87.37	86.15	73.87	78.49	75.17	78.04	82.68	82.27
	ONLY [49]	**86.00**	**86.31**	83.27	83.73	**87.47**	**86.81**	79.00	81.80	80.83	82.75	**89.47**	**89.30**	74.03	78.71	77.20	79.72	82.87	83.21
	Ours (FVR)	85.17	85.57	**83.67**	**84.34**	87.33	86.71	**79.10**	**82.05**	**81.03**	**83.15**	89.23	89.12	**74.17**	**79.06**	**78.27**	**81.17**	**84.07**	**84.43**
Adversarial	Regular	77.43	79.26	74.60	76.45	83.37	81.57	67.23	73.70	69.87	74.54	80.37	79.68	68.60	74.84	68.60	73.10	78.77	78.56
	VCD [27]	77.17	79.47	77.20	78.49	83.73	82.38	67.40	74.21	71.00	75.45	81.90	81.57	68.83	75.43	71.00	75.14	81.17	81.07
	M3ID [16]	78.23	80.22	77.47	79.14	83.37	81.62	68.60	75.11	70.10	75.16	81.90	81.26	68.67	75.28	71.17	75.36	81.90	81.57
	ONLY [49]	79.40	81.07	80.10	81.22	**83.80**	**83.46**	68.70	75.70	72.47	76.94	82.07	83.01	69.23	75.70	71.93	75.84	81.33	81.94
	Ours (FVR)	**82.87**	**83.95**	**81.07**	**82.64**	83.70	**83.51**	**69.70**	**76.27**	**72.57**	**77.24**	**83.27**	**83.24**	**69.33**	**75.90**	**72.03**	**76.90**	**82.47**	**83.12**

the official source code following this configuration, which we mark with † in the tables. Following [49], we assess our method on three open-source LVLMs, namely, LLaVA-1.5 [35], InstructBLIP [12], and Qwen-VL [2].

For all LVLM backbones, we adopt their default input prompting formats and parameters to ensure consistency with prior work. Following [27,49], we also incorporate adaptive plausibility constraints [29] during decoding with $\beta = 0.1$ across all tasks to further discourage implausible generations. Following [49], the intervention layer $\tilde{\ell}$ defaults to the first cross-attention layer. All the experiments were conducted on a single NVIDIA RTX A6000 GPU.

5.2 Comparison with State-of-the-Art

Results on POPE. Following [49], we evaluate each method under three settings derived from POPE, namely, querying random MS COCO images, querying popular MS COCO images, and an adversarial setting where questions are designed to induce mistakes. An extended evaluation that combines A-OKVQA and GQA samples is also evaluated following [31]. Table 1 reports the results. The proposed FVR consistently achieves the highest accuracy and F1 score across most setups and backbone models. For instance, on the A-OKVQA (Adversarial) subset with Qwen-VL, our method attains an accuracy of 83.27%, outperforming the second-best method by 1.20%. Even under the challenging Adversarial scenario with LLaVa-1.5 and InstructBLIP, FVR maintains a similar accuracy improvement over ONLY [49], indicating its effectiveness.

Results on CHAIR. Following its protocol, two metrics are used to evaluate each method on 500 images from MS COCO, namely, CHAIR_S (object hallucination frequency per sentence) and CHAIR_I (object hallucination per ground-truth object, measuring how often an image leads to hallucinated objects). Lower values indicate fewer hallucinations. Both the generation length settings of 64 tokens and 128 tokens are evaluated. The results in Table 2 show that, the proposed method yields the lowest hallucination rates across all three LVLM backbones. For example, with Qwen-VL and a 64-token maximum, FVR achieves

Table 2: Object hallucination rates on CHAIR [42]. We report CHAIR_S and CHAIR_I (lower is better) for captions truncated at 64 tokens and 128 tokens. Our method consistently yields the lowest hallucination frequency across models and generation lengths.

Method	LLaVA-1.5 [35]				InstructBLIP [12]				Qwen-VL [2]			
	Max Token 64		Max Token 128		Max Token 64		Max Token 128		Max Token 64		Max Token 128	
	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$
Regular	26.2	9.4	55.0	16.3	31.2	11.1	57.0	17.6	33.6	12.9	52.0	16.5
VCD [27]	24.4	7.9	54.4	16.6	30.0	10.1	60.4	17.8	33.0	12.8	50.2	16.8
M3ID [16]	21.4	6.3	56.6	15.7	30.8	10.4	62.2	18.1	32.2	11.5	49.5	17.2
Woodpecker [57]	24.9	7.5	57.6	16.7	31.2	10.8	60.8	17.6	31.1	12.3	51.8	16.3
HALC [10]	21.7	7.1	51.0	14.8	24.5	8.0	53.8	15.7	28.2	9.1	49.6	15.4
ONLY [49]	20.0	**6.2**	49.8	14.3	23.5	8.2	52.2	15.5	27.3	8.4	48.0	14.3
fuzzycd [23]†	20.8	6.5	50.4	14.7	19.8	7.9	51.3	15.2	19.4	8.2	27.6	13.2
INTER [14]	**19.0**	6.3	51.0†	15.2†	18.4	7.9	57.4†	16.9†	25.7†	11.0†	20.7†	11.5†
Ours (FVR)	19.6	**6.2**	**49.6**	**14.1**	**16.8**	**7.2**	**48.0**	**14.1**	**15.0**	**7.0**	**20.2**	**9.1**

Table 3: Results on MME-Hallucination [17] using LLaVA-1.5. Higher scores are better for each test. Our method achieves the highest scores on all sub-categories (object existence, counting, and attribute-based evaluations), yielding the best overall score.

Method	Existence	Count	Position	Color	Total
Regular	173.75	121.67	117.92	149.17	562.50
DoLa [11]	176.67	113.33	90.55	141.67	522.22
OPERA [19]	183.33	137.22	122.78	155.00	598.33
VCD [27]	186.67	125.56	128.89	139.45	580.56
M3ID [16]	186.67	128.33	131.67	151.67	598.11
Woodpecker [57]	187.50	125.00	126.66	149.17	588.33
HALC [10]	183.33	133.33	107.92	155.00	579.58
ONLY [49]	191.67	145.55	136.66	161.66	635.55
fuzzycd [23]	195	**163.33**	123.33	160	641.66
INTER [14] †	184.32	126.19	131.83	150.26	592.6
ICT [8] †	188.74	130.07	129.86	151.45	600.12
Ours (FVR)	**196.26**	142.71	**139.39**	**167.39**	**645.75**

$\text{CHAIR}_S = 15.0$ and $\text{CHAIR}_I = 7.0$, substantially lower than the second-best with 4.4 and 1.2, respectively. Consistent improvements are observed on InstructBLIP and LLaVA-1.5. These outcomes indicate that the proposed method helps the model refrain from introducing nonexistent objects in its descriptions.

Results on MME-Hallucination. Following its evaluation protocol, all the experiments use LLaVA-1.5 as the backbone. Table 3 reports the scores for object existence, object count, attribute position, and attribute color hallucination tests, respectively. Higher scores indicate better performance. The proposed method yields a total MME score of 645.75, surpassing the second-best by 4.09 points. In particular, on the Color subset, it scores 167.39, which is 5.73 points higher than the second-best. Similarly, on the Position subset, FVR reaches 139.39, 2.73 points higher than the second-best.

5.3 Ablation Studies

All the experiments are conducted on both POPE and CHAIR benchmarks. Specifically, the experiments on POPE use all 9,000 samples from MS-COCO.

Effect of Individual Components. The proposed method comprises two main components, namely, the FVR metric that guides the model's attention to verified image evidence; and the Flow-Reliability Decoding that adjusts or filters

Table 4: Ablation of each component on POPE and CHAIR. w/o FVR: without Flow Variance Reliability, w/o FRD: without Flow-Reliability Decoding.

FVR Variant	POPE				CHAIR	
	POPE Acc. ↑	Prec.	Rec.	F1	$CHAIR_S$ ↓	$CHAIR_I$ ↓
Full FVR	**85.97**	**83.55**	**89.58**	**86.39**	**19.6**	**6.2**
w/o FVR	84.08	82.06	87.17	84.52	23.5	7.8
w/o FRD	83.86	81.72	87.23	84.38	24.0	8.1
Baseline	80.42	78.20	84.59	81.27	26.2	9.4

Table 5: Comparison of different intervention strategies on POPE and CHAIR.

Strategy	POPE Acc. ↑	Prec.	Rec.	F1	$CHAIR_S$ ↓	$CHAIR_I$ ↓
Regular decoding	80.42	78.20	84.59	81.27	26.2	9.4
Zero-out $a^V_{\ell,i}$	84.26	82.13	87.69	84.82	21.2	6.9
Add noise to $a^V_{\ell,i}$	83.95	81.67	88.16	84.79	22.1	7.6
Double all $a^T_{\ell,i}$	84.37	82.52	87.55	84.96	21.6	6.8
Select heads by $\frac{\sum a^T}{\sum a^V}$	84.20	81.57	87.56	84.46	23.1	8.2
ONLY	84.91	82.84	88.07	85.37	20.0	**6.2**
FVR (Ours)	**85.97**	**83.55**	**89.58**	**86.39**	**19.6**	**6.2**

the model's output based on the grounded information. Table 4 breaks down the impact of each component. As expected, the full model performs best, indicating fewest hallucinations. Ablating the FVR guided head selection ("w/o FVR") leads to a drop in POPE accuracy (down to 84.08%) and a higher hallucination rate ($CHAIR_S$ rises to 23.5). Ablating the Flow-Reliability Decoding ("w/o FRD"), POPE accuracy falls to 83.86% and $CHAIR_S$ increases to 24.0. In this case, lacking the final check allows some hallucinated phrases to slip through. These results validate that both components of FVR are necessary.

Effect of Alternative Strategies. Table 5 summarizes some comparisons. Setting all visual attention $a^V_{\ell,i}$ to zero or adding noise to visual features yields only moderate gains in POPE accuracy and modest reductions in $CHAIR_S$. Doubling the textual attention $a^T_{\ell,i}$ had a similarly limited effect. Selecting attention heads based on the ratio of text-to-visual attention also underperforms, leading to a 1.77% drop in POPE accuracy and a 3.5 higher $CHAIR_S$. These results further show the superiority of the proposed method, which enhances the influence of visual information while diminishing unsupported signals.

Impact of Ratio K. The proposed Flow Variance Selection component selects the top K% heads measured by FVR that may contribute to the flow leakage of the visual information. Table 6 studies how K impacts the overall performance on both POPE and CHAIR. We observe that a K value of 20% achieves the optimal performance. A too small K may result in insufficient discrimination between the flow paths that contribute or do not contribute to the visual information leakage. In contrast, a too large K may mask out some paths that contribute to the final logit prediction and therefore degrade the performance.

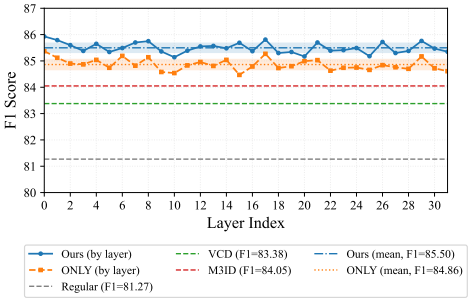


Fig. 6: FVR outperforms all the compared methods when integrated into each of the 32 transformer layers of LLaVA-1.5, indicating its effectiveness and stability. Results of Regular, VCD and M3ID are cited from [49]. The per-layer outcome of ONLY [49] is re-implemented by the official code.

Effect of Distance Metric. In the decoding stage, the proposed method uses the scaled Hellinger discrepancy basing on a scaled Hellinger gap. To validate if it is the optimal, we compare with four simpler discrepancy metrics, namely, l-1,

Table 6: Impact of the K value on POPE and CHAIR results.

K%	POPE Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	$CHAIR_S$ ↓	$CHAIR_I$ ↓
10%	85.05	82.90	87.40	85.05	20.4	6.7
20%	**85.97**	**83.55**	**89.58**	**86.39**	**19.6**	**6.2**
30%	85.92	83.45	89.40	86.32	19.9	6.4
40%	85.55	83.10	88.90	85.93	20.1	6.5

Table 7: Effect of the distance metric on POPE and CHAIR results.

Metric	POPE Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	$CHAIR_S$ ↓	$CHAIR_I$ ↓
l-1	85.23	82.94	88.21	85.49	21.5	6.7
Smooth l-1	85.11	82.72	87.96	85.27	21.7	6.8
l-2	85.62	83.18	88.93	85.92	20.4	6.5
K-L	85.79	83.36	89.12	86.12	20.1	6.4
Hellinger (Ours)	**85.97**	**83.55**	**89.58**	**86.39**	**19.6**	**6.2**

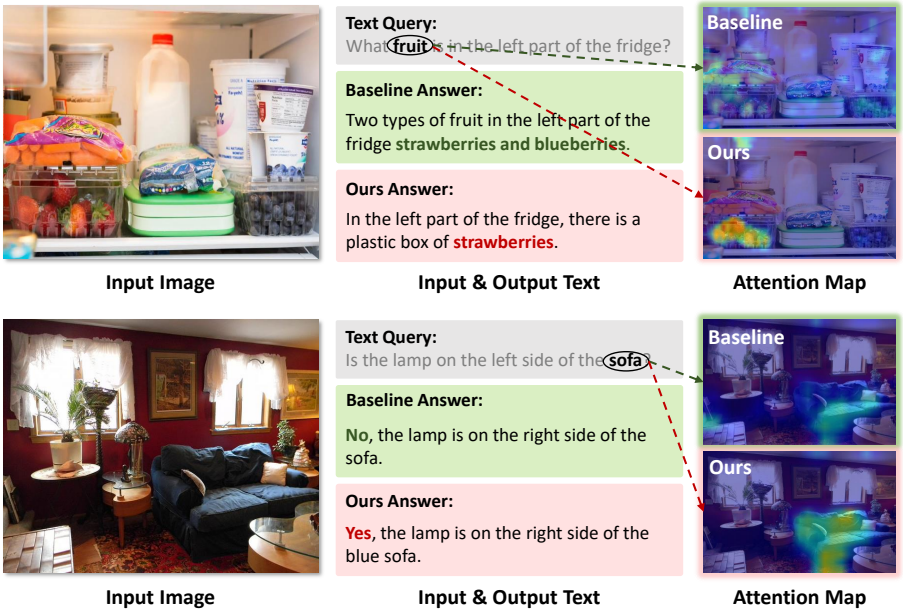


Fig. 7: Qualitative examples from LLaVA-Bench. We compare answers produced by regular decoding (baseline) and our FVR method using LLaVA-1.5. We also show token attention maps [30] of the key words in query. Hallucinated content is marked in teal, while accurate content is in magenta. In these examples, visual flow leakage to irrelevant regions leads the baseline to "misread" the object's spatial relations. In contrast, our method focuses on the query-relevant key objects, yielding accurate comprehension.

smooth l-1, l-2 and K-L divergence. Table 7 shows that the proposed scaled Hellinger discrepancy clearly outperforms the rest four alternatives on both benchmarks, indicating its effectiveness for stable flow-reliability decoding.

Robustness to Integration Layer. We further examine the robustness of the proposed FVR when embedded into different layers in the LLaVA-1.5 base model. The experiments are conducted on the full 9,000 samples of POPE MS COCO. As in [49] the per-layer numerical performance is not reported, we re-implement the experiments based on the official code. The results in Fig. 6 show that the proposed method outperforms all the compared methods on each of the 32 transformer layers, indicating its superiority, robustness and stability.

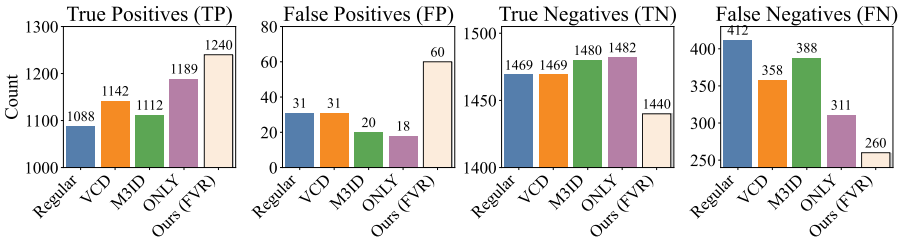


Fig. 8: Statistics of true/false positives and negatives.

5.4 Further Analysis

Case Study. We present a qualitative example from LLaVA-Bench to illustrate how FVR mitigates hallucination. For the same image and question, Fig. 7 shows the response generated by the baseline LLaVA-1.5 model (using regular greedy decoding) and the response generated by our method. While the baseline's answer includes a hallucinated object, our method's answer remains faithful to the visual content. The visualization of the token activation map [30] also confirms the observation. Compared with the baseline, the proposed method can more precisely activate the key objects that are grounded in the text query. While the cases in Fig. 3 show that FVR steers the model to be faithful to the *object count*, this example shows that FVR steers the model to describe only what it truly "sees" on *object existence* and *spatial position*. These examples overall demonstrate its effectiveness to eliminate the hallucinated content that the baseline would otherwise produce.

Model Behavior & Limitation. The proposed method reduces the hallucination by muting some heads with low FVR, which allows the model to focus more on positive samples that have high FVR. However, this approach can be confused on some negative samples that have higher FVR, therefore introducing more false positives. This pattern is clearly revealed in Fig. 8, where MS-COCO Random setting is conducted using Qwen-VL as the base model. The proposed method selects the most amount of true positive samples but also introduces some false positives. Both reduction in false negatives and increase in true positives are stronger effects, leading to a good overall hallucination alleviation. We note that this problem can be more intriguing if the positive and negative samples hold different distributions. The proposed FVR paves a way to measure the visual flow leakage, which may play an important role in analyzing the additive vs. subtractive hallucinations and inspiring hallucination benchmarks with higher granularity.

6 Conclusion

We revisited LVLM hallucination through Visual Flow Leakage (VFL), where attended visual tokens provide weak value-path delivery during decoding. We

proposed Flow Variance Reliability (FVR), a per-head score that measures gain-weighted visual value delivery and guides reliability-aware head intervention. The proposed method preserves the simplicity of training-free decoding while directly targeting the mechanism that fosters hallucination. The approach consistently improves grounding across standard LVLMs and benchmarks with negligible overhead. Beyond alleviating hallucinations, this work encourages future LVLM research to focus on measuring the quality of visual value paths, rather than only looking at where attention points.

Acknowledgements

This work is supported by the NWA-ORC Project HAICu (NWA 1518.22.105).

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4190–4197 (2020)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhao, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
3. Beran, R.: Minimum Hellinger distance estimates for parametric models. The annals of Statistics pp. 445–463 (1977)
4. Bi, Q., Shen, Y., Yi, J., Xia, G.S.: Adadcp: Learning an adapter with discrete cosine prior for clear-to-adverse domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12997–13008 (2025)
5. Bi, Q., Yi, J., Huang, H., Zheng, H., Zhan, H., Huang, Y., Li, Y., Wu, X., Zheng, Y.: Nightadapter: Learning a frequency adapter for generalizable night-time scene segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23838–23849 (2025)
6. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 782–791 (2021)
7. Chen, J., Yang, D., Wu, T., Jiang, Y., Hou, X., Li, M., Wang, S., Xiao, D., Li, K., Zhang, L.: Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024)
8. Chen, J., Zhang, T., Huang, S., Niu, Y., Zhang, L., Wen, L., Hu, X.: ICT: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4209–4221 (2025)
9. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens in large vision-language models. In: European Conference on Computer Vision. pp. 19–35 (2024)
10. Chen, Z., Zhao, Z., Luan, H., Yao, H., Li, B., Zhou, J.: Halc: Object hallucination reduction via adaptive focal-contrast decoding. In: International Conference on Machine Learning. pp. 7824–7846 (2024)

11. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J.R., He, P.: Dola: Decoding by contrasting layers improves factuality in large language models. In: International Conference on Learning Representations (2024)
12. Dai, W., Li, Z., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.C.H.: Instructblip: towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems **36**, 49250–49267 (2023)
13. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024)
14. Dong, X., Dong, S., Wang, J., Huang, J., Zhou, L., Sun, Z., Jing, L., Lan, J., Zhu, X., Zheng, B.: Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2534–2544 (2025)
15. Duan, J., Kong, F., Cheng, H., Diffenderfer, J., Kailkhura, B., Sun, L., Zhu, X., Shi, X., Xu, K.: Truthprint: Mitigating large vision-language models object hallucination via latent truthful-guided pre-intervention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7372–7382 (2025)
16. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14303–14312 (2024)
17. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Zhang, J., Zhang, Y., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
18. Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., Lin, M.: When attention sink emerges in language models: An empirical view. In: The Thirteenth International Conference on Learning Representations (2025)
19. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Yu, H., Liu, D., Zhang, W., Yu, A.Z.: Opera: Alleviating hallucination in multimodal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
20. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, Y., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination alignment via contrastive learning for multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27036–27046 (2024)
21. Jingjing, L., Feng, Y., Guo, Y., Huang, J., Ji, W., Bi, Q., Piao, Y., Zhang, M., Zhao, X., Chen, Q., et al.: Sam3-i: Segment anything with instructions. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 27232–27249 (2026)
22. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. In: The Thirteenth International Conference on Learning Representations (2025)
23. Kim, J., Kim, J., Kim, Y., Cho, S.B.: Fuzzy contrastive decoding to alleviate object hallucination in large vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20572–20581 (2025)
24. Kobayashi, G., Kuribayashi, T., Yokoi, S., Inui, K.: Attention is not only a weight: Analyzing transformers with vector norms. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 7057–7075 (2020)
25. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)

26. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Mitigating hallucinations in large vision-language models with instruction-contrastive decoding. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 15840–15853 (2024)
27. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: 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. pp. 13872–13882 (2024)
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742 (2023)
29. Li, X.L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T.B., Zettlemoyer, L., Lewis, M.: Contrastive decoding: Open-ended text generation as optimization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 12286–12312 (2023)
30. Li, Y., Wang, H., Ding, X., Wang, H., Li, X.: Token activation map to visually explain multimodal LLMs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 48–58 (2025)
31. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
33. Lindsay, B.G.: Efficiency versus robustness: The case for minimum Hellinger distance and related methods. The annals of statistics **22**(2), 1081–1114 (1994)
34. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
35. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
36. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, OCR, and world knowledge. arXiv preprint arXiv:2402.00525 (2024)
37. Manevich, A., Tsarfaty, R.: Mitigating hallucinations in large vision-language models (LVLMs) via language-contrastive decoding (LCD). In: Findings of the Association for Computational Linguistics ACL 2024. pp. 6008–6022 (2024)
38. Merullo, J., Eickhoff, C., Pavlick, E.: Talking heads: Understanding inter-layer communication in transformer language models. Advances in Neural Information Processing Systems **37**, 61372–61418 (2024)
39. Michel, P., Levy, O., Neubig, G.: Are sixteen heads really better than one? Advances in neural information processing systems **32** (2019)
40. Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Clip-DPO: Vision-language models as a source of preference for fixing hallucinations in LVLMs. In: European Conference on Computer Vision. pp. 395–413 (2024)
41. Peng, S., Yang, S., Jiang, L., Tian, Z.: Mitigating object hallucinations via sentence-level early intervention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 635–646 (2025)
42. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045 (2018)

43. Shen, Y., Bi, Q., Huang, J.H., Zhu, H., Pimentel, A.D., Pathania, A.: Macp: Minimal yet mighty adaptation via hierarchical cosine projection. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 20602–20618 (2025)
44. Shen, Y., Bi, Q., Huang, J.h., Zhu, H., Pimentel, A.D., Pathania, A.: Ssh: Sparse spectrum adaptation via discrete hartley transformation. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 10400–10415 (2025)
45. Shen, Y., Bi, Q., Wang, Z., Yang, Z., Wang, C., Zhang, Z., Tiwari, P., Pimentel, A.D., Pathania, A.: Efficient multimodal spatial reasoning via dynamic and asymmetric routing. In: The Fourteenth International Conference on Learning Representations (2026)
46. Skean, O., Arefin, M.R., Zhao, D., Patel, N.N., Naghiyev, J., LeCun, Y., Shwartz-Ziv, R.: Layer by layer: Uncovering hidden representations in language models. In: Forty-second International Conference on Machine Learning (2025)
47. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented RLHF. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 13088–13110 (2024)
48. Wan, Z., Xie, Y., Zhang, C., Lin, Z., Wang, Z., Stepputtis, S., Ramanan, D., Sycara, K., Xie, Y.: Instruct-part: Task-oriented part segmentation with instruction reasoning. arXiv preprint arXiv:2505.18291 (2025), preprint
49. Wan, Z., Zhang, C., Yong, S., Ma, M.Q., Stepputtis, S., Morency, L.P., Ramanan, D., Sycara, K., Xie, Y.: Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3225–3234 (2025)
50. Wang, S., Zhang, Y., Zhu, Y., Liu, E., Li, J., Liu, Y., Ji, X.: Shift: Smoothing hallucinations by information flow tuning for multimodal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3639–3649 (2025)
51. Wang, X., Pan, J., Ding, L., Biemann, C.: Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In: Findings of the Association for Computational Linguistics 2024. pp. 15840–15853 (2024)
52. Wei, Z., Chen, L., Jin, Y., Ma, X., Liu, T., Ling, P., Wang, B., Chen, H., Zheng, J.: Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28619–28630 (2024)
53. Wu, Y., Wang, Y., Tang, S., Wu, W., He, T., Ouyang, W., Torr, P., Wu, J.: Dettoolchain: A new prompting paradigm to unleash detection ability of LLM. In: European Conference on Computer Vision. pp. 164–182 (2024)
54. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: The Twelfth International Conference on Learning Representations (2024)
55. Yang, Z., Luo, X., Han, D., Xu, Y., Li, D.: Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10610–10620 (2025)
56. Yi, J., Bi, Q., Zheng, H., Zhan, H., Ji, W., Huang, Y., Li, Y., Zheng, Y.: Learning spectral-decomposited tokens for domain generalized semantic segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8159–8168 (2024)

57. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences **67**(12), 220105 (2024)
58. Zeng, Y., Li, W., Han, J., Zhou, K., Loy, C.C.: Contextual object detection with multimodal large language models. International Journal of Computer Vision pp. 1–29 (2024)
59. Zheng, G., Qian, J., Tang, J., Yang, S.: Why LVLMs are more prone to hallucinations in longer responses: The role of context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4101–4113 (2025)
60. Zhi, Z., Jin, P., Xiao, Y., He, T., Han, Z., Zhang, Z., Hou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.13900 (2024)