

Visual Representation Regularization for Multimodal Hallucination Mitigation

Yifan Meng^{1†}, Enguang Wang^{1†}, Xusheng Cao¹, Fei Yang^{2,1}, Xialei Liu^{1*}

¹VCIP, CS, Nankai University ²NKIARI, Shenzhen Futian

{mengyifan, enguangwang, caoxusheng}@mail.nankai.edu.cn, {feiyang, xialei}@nankai.edu.cn

Abstract

Multimodal large language models (MLLMs) have achieved remarkable progress in multimodal reasoning, yet they still suffer from hallucinations that undermine their reliability. Existing inference-time mitigation methods mainly focus on optimizing decoding strategies or analyzing internal attention mechanisms, while the role of visual representations remains underexplored. In this work, we reveal that visual representations exhibit information dispersion during layer-wise evolution, manifested as a progressive deviation from the initial visual subspace and increasing alignment with the textual subspace. Based on this insight, we propose Visual Representation Regularization (VRR), a training-free method that mitigates hallucinations by regularizing visual representations toward their initial states. Extensive experiments demonstrate that VRR consistently reduces hallucinations and is compatible with existing mitigation methods. Our code is available at <https://github.com/Jiarker/VRR>.

1 Introduction

Multimodal large language models (MLLMs) have shown strong capabilities in multimodal understanding and reasoning (Qin et al., 2025). However, they still suffer from hallucinations, where generated content is inconsistent with the visual input (Bai et al., 2025; Liu et al., 2024a). Such hallucinations substantially undermine the reliability and trustworthiness of MLLMs, thereby limiting their deployment in real-world applications.

Recent studies have explored inference-time hallucination mitigation from different perspectives (Huang et al., 2025). One line of work improves decoding strategies by contrasting output distributions under different visual conditions, thereby suppressing responses dominated by language priors (Zhu et al., 2025; Yang et al., 2025a;

Chen et al., 2024). Another line investigates hallucinations from the perspective of model interpretability and internal mechanisms, such as identifying visually grounded attention heads or analyzing attention patterns associated with hallucinations (Huang et al., 2024; Zhong et al., 2026). Although these methods provide valuable insights, the role of visual representations in hallucination formation remains underexplored.

Since textual responses are generated from hidden states that interact with visual tokens through LLM layers, grounded generation requires visual evidence to remain accessible and distinguishable throughout this layer-wise transformation. However, as visual tokens are processed by the LLM, visual evidence may become diluted across the representation space, making visual cues less accessible (Wang et al., 2026; Jiang et al., 2025; Neo et al., 2025). To this end, we analyze the layer-wise evolution of visual representations in MLLMs and reveal an information dispersion phenomenon: visual representations gradually spread from a compact low-dimensional structure to a more dispersed high-dimensional space. Moreover, from a geometric perspective, we find that visual representations progressively deviate from the initial visual subspace and become increasingly aligned with the textual subspace. This suggests that visual evidence may be diluted during evolution, which potentially increases the risk of hallucination.

Motivated by these observations, we propose Visual Representation Regularization (VRR), a simple yet effective training-free method to mitigate hallucinations. Specifically, VRR regularizes visual representations toward their initial states to maintain compact representations and recover the original visual information. However, direct alignment may damage the linguistic information contained in the visual representations, while also disrupting their inherent visual information structure. Therefore, VRR performs regularization within the

[†]Equal contribution.

^{*}Corresponding author.

initial visual subspace, where the dominant visual structure of the initial representation is preserved. In this way, VRR restores concentrated visual evidence with minimal modification to the evolved representations, thereby strengthening visual grounding for mitigating hallucinations without additional training or parameter updates.

Our main contributions are as follows: (1) We reveal the information dispersion of visual representations during layer-wise evolution and analyze it from a geometric perspective, providing new insights into the internal mechanism of hallucination. (2) We propose VRR, a training-free method that mitigates hallucinations by regularizing visual representations toward their initial states. (3) Extensive experiments on multiple benchmarks and MLLMs demonstrate the effectiveness, robustness, and compatibility of VRR with existing inference-time hallucination mitigation methods.

2 Related Work

2.1 Cross-Modal Transformation of Visual Representations

In recent years, analyzing foundation models through internal representation (Dang et al., 2024; Lin et al., 2025) has become an important direction in interpretability research. Visual information does not remain static after entering the model; instead, it is continuously reorganized through multi-layer Transformer propagation and interacts with language representation.

Existing studies investigate the layer-wise evolution of internal visual representations in MLLMs. (Neo et al., 2025) observes that visual representations become increasingly interpretable within the vocabulary space, manifested as the alignment between image content and its corresponding textual representations. (Wang et al., 2026) argues that visual representations are not optimized for preserving visual fidelity during layer-wise evolution, but instead undergo implicit and progressive transformations with the sole objective of better serving the final text generation task.

2.2 Mitigating Hallucinations During Inference

Hallucination in MLLMs refers to a mismatch between the generated text and the actual content in the corresponding image (Leng et al., 2024; Zhuang et al., 2024; Phukan et al., 2025; Lyu et al., 2026; Chen et al., 2024). When hallucinations occur, the

model may omit objects present in the image, describe objects that do not exist in the image, or incorrectly interpret the image content (Bai et al., 2025). Many studies focus on mitigating hallucinations during inference.

Hallucinations can be mitigated during decoding. VCD (Leng et al., 2024) analyzes the effects of statistical bias and over-reliance on unimodal priors, and mitigates hallucinations by contrasting the output distributions of clean and corrupted visual inputs. DoLa (Chuang et al., 2024) observes that factual knowledge in LLMs is localized in specific layers, and leverages inter-layer contrastive decoding to better exploit factual information.

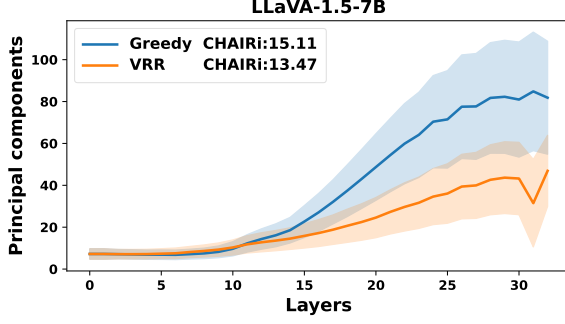
Another line of research focuses on enhancing visual information. The mainstream approach aims to modify the attention mechanism to guide the model toward visual information. VHR (He et al., 2025) mitigates hallucinations by strengthening the effects of vision-aware attention heads. EAH (Zhang et al., 2024) investigates hallucinations from the perspective of attention sinks.

Recent works have explored hallucination mitigation from the representation perspective. Nullu (Yang et al., 2025b) constructs guidance vectors from external data sources to manipulate model representations for hallucination reduction. DAMO (Wang et al., 2025) aligns representations before the decoding stage to maintain decoding consistency across layers.

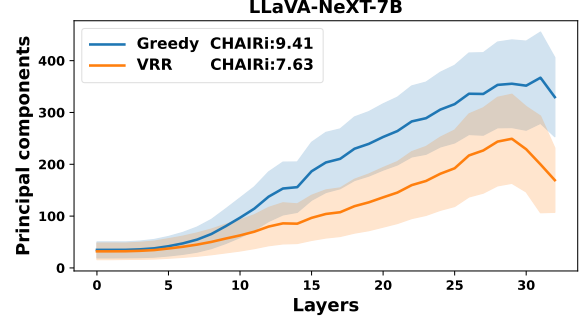
Compared to these methods, VRR directly regularizes visual representations toward their initial states at each layer, without relying on external data sources. VRR restores rich visual information, thereby mitigating hallucinations. Moreover, VRR does not require multiple forward passes and is compatible with various decoding strategies and hallucination mitigation methods.

3 Inter-layer Evolution of Visual Representations

In this section, we first introduce the formulation of MLLMs, and then analyze how visual representations evolve across LLM layers. We reveal an information dispersion phenomenon, where visual representations gradually spread from a compact low-dimensional structure to a more dispersed space. We further provide a geometric explanation by showing that visual representations progressively deviate from the initial visual subspace and align with the layer-wise textual subspace.



(a) VRR decreases CHAIR_i from 15.11 to 13.47.



(b) VRR decreases CHAIR_i from 9.41 to 7.63.

Figure 1: Layer-wise effective dimensionality r^l of visual representations. VRR reduces representation dispersion, thereby preserving compact visual representations and mitigating the risk of hallucination.

3.1 Preliminaries

The Architecture of MLLMs. MLLMs typically integrate a pretrained vision encoder with a large language model. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$, the vision encoder extracts a sequence of N_v patch features. A projector then maps these features into the language embedding space, producing the projected vision tokens $H_v^0 = \{h_{v,1}^0, \dots, h_{v,N_v}^0\}$, where $h_{v,i}^0 \in \mathbb{R}^d$ and H_v^0 denotes the initial visual representation.

Given a text prompt, the MLLM generates response tokens autoregressively. At each layer l , we divide the hidden states into visual representations H_v^l and textual representations H_t^l . The textual representation includes both prompt tokens and generated response tokens:

$$H_t^l = \{h_{t,1}^l, \dots, h_{t,N_t}^l\}, \quad N_t = N_p + N_g, \quad (1)$$

where N_p and N_g denote the numbers of prompt tokens and generated tokens, respectively.

3.2 Information Dispersion

Hallucinations in MLLMs are closely related to insufficient visual grounding during response generation (He et al., 2025; Li et al., 2025; Zhuang et al., 2024). Since textual responses are generated from hidden states that interact with visual tokens through the LLM layers, grounded generation requires visual evidence to remain accessible and distinguishable throughout this layer-wise transformation. However, as visual tokens are repeatedly transformed by the LLM, visual evidence may become diluted across the representation space, making visual cues less accessible and increasing reliance on language priors (Wang et al., 2026; Jiang et al., 2025; Neo et al., 2025).

To characterize this dilution, we analyze how visual information is distributed in the representation space across layers. Specifically, we perform Singular Value Decomposition (SVD) on the visual representation H_v^l at each layer l and compute the number of principal components r^l required to preserve 95% of its spectral energy. Thus, r^l can be the effective dimensionality of H_v^l . A smaller r^l indicates that dominant visual information is concentrated in a compact low-dimensional subspace, making visual evidence easier to preserve and access. In contrast, a larger r^l suggests that visual information is distributed across more dimensions, indicating stronger information dispersion.

We conduct this analysis on LLaVA-1.5-7B (Liu et al., 2023) and LLaVA-NeXT-7B (Liu et al., 2024c) using the COCO dataset (Lin et al., 2014). Given each image, we prompt the model with “Please describe the image in detail” and extract the layer-wise visual representations during generation. For each sample and each layer, we compute r^l and report its average over all samples.

As shown in Figure 1, the visual representations show a clear increase in r^l as the layer increases, showing that visual representations gradually shift from a compact low-dimensional structure to a more dispersed high-dimensional space. This suggests that visual evidence becomes increasingly scattered across the representation space during layer-wise transformation. Such dispersion may weaken visual grounding by making visual cues less accessible to the generation process, which potentially increases the risk of hallucination.

3.3 Geometric Explanation

To further understand why visual information becomes increasingly dispersed across layers, we examine the evolution of visual representations from

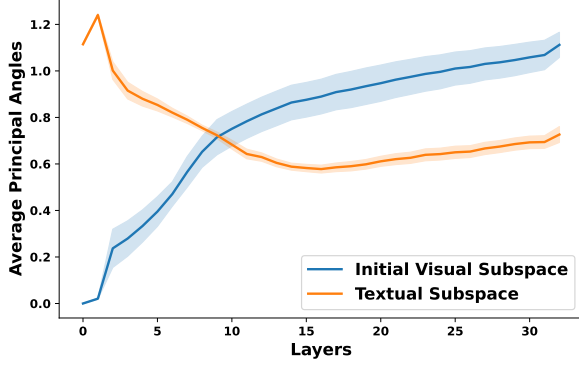


Figure 2: Layer-wise average principal angles for visual-subspace evolution on LLaVA-1.5-7B.

a geometric perspective. Specifically, we compare the visual subspace at each layer with two reference subspaces: the initial visual subspace, which preserves the compact visual structure before LLM processing, and the layer-wise textual subspace, which characterizes the language-oriented geometry introduced by the LLM.

Initial Visual Subspace P_v^0 . We first construct the initial visual subspace from the projected visual representation H_v^0 , which serves as a reference for the original visual structure. Specifically, we perform SVD on H_v^0 and retain the principal right singular vectors that preserve 95% of its spectral energy, denoted as V_v^0 . The corresponding projection matrix is defined as

$$P_v^0 = V_v^0 (V_v^0)^\top. \quad (2)$$

Since P_v^0 is constructed from the dominant directions of H_v^0 , it captures the compact visual structure of the initial visual tokens. Meanwhile, as in Section 3.2, H_v^0 exhibits a low effective dimensionality, making P_v^0 a suitable reference for analyzing how visual representations deviate from their original geometry across LLM layers.

Textual Subspace P_t^l . We further construct a layer-wise textual subspace to characterize the linguistic information introduced by the LLM. Since the text tokens of a single sample are limited and may contain image-conditioned information due to cross-modal self-attention, directly constructing a subspace from each individual sample can be unstable. Therefore, we aggregate text representations from multiple samples to obtain a more reliable empirical approximation of the text space.

Specifically, inspired by (Yang et al., 2025b), we randomly sample M instances from the COCO 2014 training set and collect their text representa-

tions at each layer:

$$\tilde{H}_t^l = \{H_t^{l,(m)}\}_{m=1}^M, \quad (3)$$

where $H_t^{l,(m)}$ denotes the text hidden states of the m -th sample at layer l . Following the construction of P_v^0 , we perform SVD on $\tilde{H}_t^l$ and retain the principal right singular directions to obtain the projection matrix P_t^l . The resulting P_t^l serves as an empirical layer-wise textual subspace, providing a reference for measuring whether visual representations become increasingly aligned with the language-oriented geometry of the LLM.

Average Principal Angles. Following the construction of P_v^0 , we construct the visual subspace P_v^l from H_v^l at each layer. To quantify its geometric evolution, we compute the average principal angles (Knyazev and Argentati, 2002) between P_v^l and two reference subspaces: the initial visual subspace P_v^0 and the textual subspace P_t^l . These angles are denoted as $\theta_{v,0}^l$ and $\theta_{v,t}^l$, respectively. A smaller angle indicates stronger subspace overlap and closer geometric alignment.

Figure 2 presents the layer-wise evolution of $\theta_{v,0}^l$ and $\theta_{v,t}^l$. As the layer depth increases, $\theta_{v,0}^l$ gradually becomes larger, indicating that the visual subspace progressively deviates from the initial visual subspace. Meanwhile, $\theta_{v,t}^l$ consistently decreases, suggesting that the visual subspace becomes increasingly aligned with the textual subspace. These trends reveal a clear geometric shift during layer-wise evolution: visual representations gradually move away from their original compact visual geometry and become more coupled with the language-oriented geometry of the LLM.

Such a shift can help incorporate linguistic priors into visual representations, which is beneficial for multimodal reasoning (Jiang et al., 2025). However, excessive alignment with the textual subspace may also reshape the original visual geometry and scatter visual cues that are initially concentrated in H_v^0 . This provides a geometric explanation for the information dispersion observed in Section 3.2: as visual representations deviate from the compact initial visual subspace and become increasingly aligned with the textual subspace, visual evidence may become diluted, weakening visual grounding during response generation.

4 Visual Representation Regularization

The above analysis shows that visual representations become increasingly dispersed as they propa-

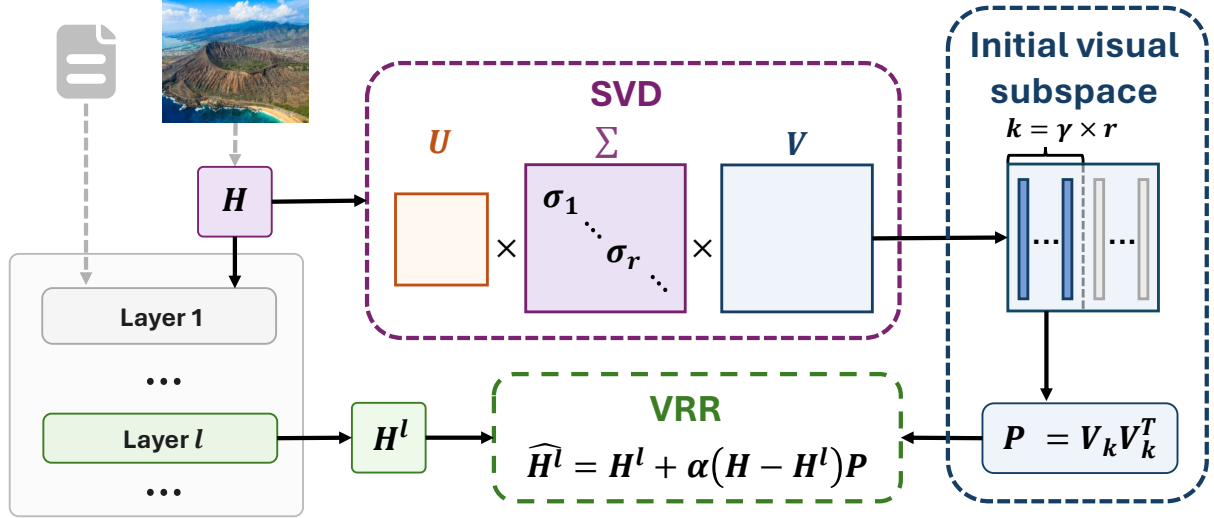


Figure 3: Visualization of SVD, initial visual subspace construction and VRR. VRR performs complete regularization by pulling H^l toward H in the initial visual subspace.

gate through LLM layers, making visual evidence less accessible for grounded generation. To mitigate this issue, we propose Visual Representation Regularization (VRR), a training-free method that performs subspace-aware correction on the layer-wise visual representation. VRR leverages the compact structure preserved in the initial visual representation H_v^0 to regularize H_v^l , while keeping the updated representation close to its original layer-wise state. In this way, VRR aims to recover concentrated visual evidence without overwriting the language-aligned information accumulated through LLM layers, as illustrated in Figure 3.

Directly forcing H_v^l to match H_v^0 may disrupt the multimodal information formed during layer-wise evolution. Therefore, instead of global alignment, VRR only pulls H_v^l toward H_v^0 within a selected initial visual subspace. For notational simplicity, we denote H_v^0 as H . We perform SVD on H :

$$H = U\Sigma V^\top, \quad (4)$$

where U and V are the left and right singular matrices, and Σ contains the singular values. Assume that there are r non-zero singular values sorted in descending order:

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0. \quad (5)$$

We select the top $k = \lfloor \gamma \times r \rfloor$ right singular vectors, denoted as V_k , where $\gamma \in (0, 1]$ controls the size of the regularization subspace. The corresponding projection matrix is defined as

$$P = V_k V_k^\top. \quad (6)$$

For simplicity, we denote H_v^l as H^l and let $\hat{H}^l$ be the regularized visual representation at layer l . VRR is formulated as the following subspace-aware optimization problem:

$$\mathcal{J}(\hat{H}^l) = \|\hat{H}^l - H^l\|_F^2 + \frac{\alpha}{1 - \alpha} \|(\hat{H}^l - H)P\|_F^2, \quad (7)$$

where $\alpha \in [0, 1)$ controls the strength of the visual regularization. The first term keeps $\hat{H}^l$ close to the evolved representation H^l , preventing the correction from overwriting the information accumulated through LLM layers. The second term encourages $\hat{H}^l$ to recover the compact visual structure of H within the selected initial visual subspace.

Minimizing Eq. (7) yields the following closed-form solution:

$$\hat{H}^l = H^l + \alpha(H - H^l)P. \quad (8)$$

The detailed derivation is provided in Appendix A.

It can be interpreted as a residual correction applied to H^l . $(H - H^l)P$ pulls the dispersed visual representation back toward the compact visual structure of H , while the residual form preserves the original layer-wise representation. Therefore, VRR mitigates information dispersion in a training-free manner and provides more accessible visual evidence for grounded generation.

5 Experiments

5.1 Benchmarks.

We evaluate our method on CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023), MME (Fu

Table 1: Experimental results of CHAIR evaluation. Values in parentheses denote the performance changes, where **green** indicates improvement and **yellow** indicates degradation.

Baselines	mPLUG-Owl2-7B		LLaVA-1.5-7B		LLaVA-NeXT-7B	
	CHAIR _S ↓	CHAIR _i ↓	CHAIR _S ↓	CHAIR _i ↓	CHAIR _S ↓	CHAIR _i ↓
Greedy	57.72	16.90	49.68	14.32	29.08	8.08
w/VRR	40.40 (17.32)	12.12 (4.78)	44.60 (5.08)	13.01 (1.31)	23.80 (5.28)	6.75 (1.33)
Beam	56.48	16.28	53.84	15.60	25.72	6.92
w/VRR	43.08 (13.40)	12.52 (3.76)	47.04 (6.80)	13.90 (1.70)	24.24 (1.48)	6.63 (0.29)
VCD	67.92	18.74	51.92	15.42	30.80	8.72
w/VRR	44.80 (23.12)	12.73 (6.01)	45.52 (6.40)	13.34 (2.08)	27.16 (3.64)	7.73 (0.99)
VHR	58.80	15.74	33.32	9.71	24.96	6.80
w/VRR	36.20 (22.60)	10.52 (5.22)	30.28 (3.04)	9.26 (0.45)	23.00 (1.96)	6.01 (0.79)
Nullu	50.88	16.58	44.40	13.21	34.04	8.59
w/VRR	33.56 (17.32)	11.87 (4.71)	39.24 (5.16)	11.66 (1.55)	30.24 (3.80)	7.96 (0.63)

et al., 2026) and LLaVA-Bench (In-the-Wild) (Liu et al., 2024b). CHAIR measures object hallucination in image captioning, and we report caption-level (CHAIR_S) and instance-level (CHAIR_i) hallucination rates. POPE evaluates object-level hallucination with binary object-existence questions. Following the standard setting, we use COCO (Lin et al., 2014), GQA (Hudson and Manning, 2019), and A-OKVQA (Schwenk et al., 2022), each with random, popular, and adversarial subsets, and report average accuracy and F1 score. MME is a comprehensive benchmark comprising 14 tasks across two categories: Perception and Cognition. We report the results on the Count, Position, and OCR tasks, as well as the overall scores for the two categories. LLaVA-Bench is designed to evaluate vision-language models on a broad range of challenging tasks, including simple question answering, detailed description, and complex reasoning. Due to the open-ended and complex nature of the responses, GPT-4V is used to evaluate LVLM outputs in terms of accuracy, detailedness, and naturalness. More details are provided in Appendix B.1.

5.2 Experimental Settings

This section introduces the MLLMs and baselines used in our experiments. More details are provided in Appendix B.2.

Implementation Details. We conduct experiments on four representative MLLMs: LLaVA-1.5-7B (Liu et al., 2023), mPLUG-Owl2-7B (Ye et al., 2024), LLaVA-NeXT-7B (Liu et al., 2024c) and LLaVA-NeXT-13B. For mPLUG-Owl2-7B, we use $\gamma = 0.5, \alpha = 0.05$. For LLaVA-1.5-7B, we use

$\gamma = 0.5, \alpha = 0.4$. For LLaVA-NeXT-7B, we use $\gamma = 0.2, \alpha = 0.2$. We conduct broader experiments using LLaVA-NeXT-13B with $\gamma = 0.2$ and $\alpha = 0.2$. We apply VRR to all LLM layers by default. Since applying VRR to the middle and later layers degrades LLaVA-NeXT-13B performance, we apply VRR only to the first five layers. We use $M = 1000$ samples to construct the textual subspace in Section 3.3. The maximum number of generated tokens is set to 512 for LLaVA-1.5-7B and LLaVA-NeXT-7B, while it is set to 128 for mPLUG-Owl2-7B. For beam decoding, the number of beams is set to 5. The default decoding strategy is greedy decoding.

Baselines. VRR is a plug-and-play method that is compatible with popular training-free methods. We use five baselines: greedy sampling, beam sampling, VCD (Leng et al., 2024), VHR (He et al., 2025) and Nullu (Yang et al., 2025b). Greedy sampling and beam sampling are two commonly used decoding strategies. VCD and VHR mitigate hallucinations from contrastive decoding and the attention mechanism, respectively. Both our method VRR and the baseline Nullu are representation-based approaches for hallucination mitigation.

5.3 Results

This section introduces the main results in our experiments. More details are provided in Appendix B.3.

CHAIR. As shown in Table 1, VRR consistently mitigates hallucinations under most settings, demonstrating its effectiveness and robust-

Table 2: Experimental results of POPE evaluation. Values in parentheses denote the performance changes, where **green** indicates improvement and **yellow** indicates degradation.

Datasets	Baselines	mPLUG-Owl2-7B		LLaVA-1.5-7B		LLaVA-NeXT-7B	
		Acc $\uparrow$	F1 $\uparrow$	Acc $\uparrow$	F1 $\uparrow$	Acc $\uparrow$	F1 $\uparrow$
MSCOCO	Greedy	78.14	81.04	83.72	84.77	88.67	88.48
	w/VRR	83.17 (5.03)	84.31 (3.27)	84.79 (1.07)	85.58 (0.81)	88.98 (0.31)	88.69 (0.21)
	Beam	78.17	81.07	84.81	85.12	88.63	88.44
	w/VRR	83.15 (4.98)	84.29 (3.22)	84.70 (0.11)	85.51 (0.39)	88.96 (0.33)	88.67 (0.23)
	VCD	77.98	80.64	83.04	84.26	88.34	88.11
	w/VRR	82.13 (4.15)	83.23 (2.59)	83.71 (0.67)	84.71 (0.45)	88.54 (0.20)	88.23 (0.12)
GQA	VHR	79.97	82.21	83.80	85.25	89.25	89.02
	w/VRR	84.58 (4.61)	85.23 (3.02)	84.52 (0.72)	85.62 (0.37)	89.00 (0.25)	88.59 (0.43)
	Nullu	83.37	84.28	84.80	85.31	88.72	88.08
	w/VRR	85.91 (2.54)	85.95 (1.67)	85.69 (0.89)	85.88 (0.57)	88.71 (0.01)	87.95 (0.13)
	Greedy	75.59	79.66	74.41	79.49	83.00	84.94
	w/VRR	80.19 (4.60)	81.87 (2.21)	75.84 (1.43)	80.28 (0.79)	83.57 (0.57)	85.16 (0.22)
A-OKVQA	Beam	75.48	79.58	74.35	79.46	83.09	85.01
	w/VRR	80.12 (4.64)	81.82 (2.24)	75.74 (1.39)	80.22 (0.76)	83.79 (0.70)	85.34 (0.33)
	VCD	76.15	79.44	73.50	78.81	82.81	84.74
	w/VRR	79.88 (3.73)	80.76 (1.32)	74.44 (0.94)	79.38 (0.57)	83.27 (0.46)	84.91 (0.17)
	VHR	75.47	79.44	74.10	79.33	83.91	85.42
	w/VRR	80.69 (5.22)	81.73 (2.29)	74.32 (0.22)	79.30 (0.03)	84.36 (0.45)	85.45 (0.03)
A-OKVQA	Nullu	80.69	81.99	77.53	81.33	85.13	86.23
	w/VRR	81.82 (1.13)	82.21 (0.22)	78.81 (1.28)	81.96 (0.63)	85.23 (0.10)	85.89 (0.34)
	Greedy	71.95	77.83	76.04	80.30	85.47	87.02
	w/VRR	78.27 (6.32)	81.51 (3.68)	77.33 (1.29)	80.99 (0.69)	85.91 (0.44)	87.20 (0.18)
	Beam	71.89	77.80	75.99	80.26	85.47	87.02
	w/VRR	78.22 (6.33)	81.47 (3.67)	77.26 (1.27)	80.94 (0.68)	85.96 (0.49)	87.24 (0.22)
A-OKVQA	VCD	71.66	77.43	75.23	79.66	85.28	86.84
	w/VRR	77.44 (5.78)	80.62 (3.19)	76.34 (1.11)	80.34 (0.68)	85.76 (0.48)	87.11 (0.27)
	VHR	72.91	78.38	75.72	80.13	86.20	87.49
	w/VRR	78.98 (6.07)	81.76 (3.38)	76.34 (0.62)	80.29 (0.16)	86.46 (0.26)	87.48 (0.01)
	Nullu	80.40	82.71	78.59	81.82	86.81	87.76
	w/VRR	82.33 (1.93)	83.72 (1.01)	80.18 (1.59)	82.65 (0.83)	86.81 (0.00)	87.46 (0.30)

ness. VRR brings the most significant improvement to mPLUG-Owl2-7B. Taking greedy decoding as an example, CHAIR_S and CHAIR_i are reduced by 17.32 and 4.78, respectively, in contrast to the marginal gains achieved by other methods. For LLaVA-1.5-7B, VRR yields consistent improvements. Although LLaVA-NeXT-7B already achieves strong performance, VRR further improves its results. For example, under VHR, adding VRR further reduces CHAIR_S and CHAIR_i by 1.96 and 0.79, respectively. Results also prove that VRR is compatible with different decoding strategies and other hallucination mitigation methods.

POPE. Experimental results of POPE evaluation are provided in Table 2. Consistent with the results on the CHAIR benchmark, VRR brings the most significant improvement to mPLUG-Owl2-7B. For example, under greedy decoding on A-OKVQA, VRR improves Acc by 6.32 and F1 by 3.68. VRR also yields consistent gains on LLaVA-1.5-7B across the three datasets. Although LLaVA-NeXT-7B already substantially reduces hallucinations compared with other MLLMs, VRR still brings positive effects in most setting. For instance, on GQA with VHR, where the Acc already reaches 83.91, VRR further improves it by 0.45. VRR consistently improves performance, demonstrating

Table 3: Experimental results on MME and LLaVA-Bench benchmarks, with the best results highlighted in **bold**.

Model	MME					LLaVA-Bench		
	Count	Position	OCR	Perception	Cognition	Accuracy	Detailedness	Naturalness
LLaVA-NeXT-7B w/ VRR	133.33 135.00	151.67 140.00	125.0 130.0	1433.01 1441.02	192.50 207.14	5.517 5.533	6.417 6.683	6.783 7.133
LLaVA-NeXT-13B w/ VRR	151.67 153.33	125.00 130.00	90.0 97.5	1433.93 1444.34	273.21 278.21	5.300 5.317	6.767 6.817	6.333 6.450

strong plug-and-play capability.

MME and LLaVA-Bench. To broaden the evaluation, we evaluate VRR on MME and LLaVA-Bench using both LLaVA-NeXT-7B and the larger LLaVA-NeXT-13B model. The results are summarized in Table 3. VRR remains competitive on the newly added benchmarks. On MME, VRR improves performance on most evaluation tasks, suggesting that it can mitigate object-level hallucinations while preserving broader visual perception and reasoning capabilities. On LLaVA-Bench, all evaluated models consistently benefit from VRR, demonstrating that it mitigates hallucinations without compromising the overall multimodal capability of MLLMs.

5.4 Further Analysis

Following (An et al., 2025), we conduct ablation studies on the popular subset of the POPE benchmark constructed from the COCO dataset and provide further analysis. More details are provided in Appendix B.4.

Table 4: Ablation experiments of α conducted on the three models, with the best results highlighted in **bold**.

α	mPLUG-Owl2-7B		LLaVA-1.5-7B		LLaVA-NeXT-7B	
	Acc $\uparrow$	F1 $\uparrow$	Acc $\uparrow$	F1 $\uparrow$	Acc $\uparrow$	F1 $\uparrow$
0.0	77.40	80.43	84.80	85.50	88.97	88.70
0.05	83.13	84.19	85.10	85.80	89.23	88.93
0.1	86.43	86.40	85.40	86.08	89.40	89.06
0.2	50.00	66.67	85.57	86.17	89.60	89.22
0.4	50.00	66.67	86.00	86.45	89.70	89.22
0.6	50.00	66.67	86.13	86.49	89.37	88.80

Ablation Experiments of α . Table 4 presents the ablation results of α , which controls the regularization strength in VRR. The optimal value of α differs across models. mPLUG-Owl2-7B performs well at $\alpha = 0.05$ and achieves the best performance at $\alpha = 0.1$. However, applying $\alpha = 0.1$ to other baselines leads to performance degradation, indicating that this setting is less robust. Moreover,

when a large α is used, the model performance becomes comparable to random selection, because the output format is disrupted and the model fails to provide standardized answers. For LLaVA-1.5-7B, the performance steadily improves as α increases. For LLaVA-NeXT-7B, increasing α from 0.0 to 0.4 brings substantial performance gains. The Acc values at $\alpha = 0.2$ and $\alpha = 0.4$ are close, and their F1 scores are the same. To improve robustness across different settings and datasets, we finally set α to 0.05, 0.4, and 0.2 for mPLUG-Owl2-7B, LLaVA-1.5-7B, and LLaVA-NeXT-7B, respectively.

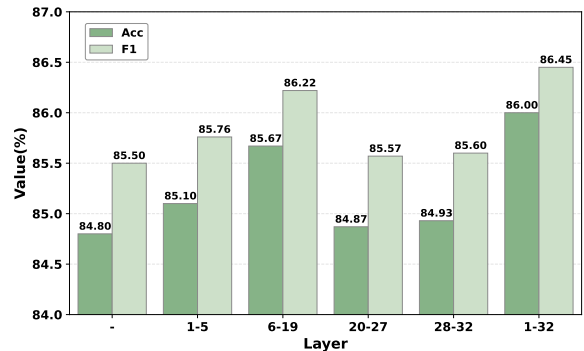


Figure 4: Ablation experiments on layer.

Layer Selection. We investigate the effect of layer selection on VRR by conducting experiments on LLaVA-1.5-7B. The results are shown in Figure 4. “-” indicates the setting without VRR. Following (Jiang et al., 2025), we divide the LLM layers into shallow layers (1–5), middle layers (6–19), middle-to-late layers (20–27), and late layers (28–32), and compare them with all layers (1–32). Our method improves model performance under all settings, with the most significant gains observed in the middle layers. Applying VRR to all layers achieves the best overall performance.

Visual Subspace Construction. VRR performs regularization within the initial visual subspace. To validate this design, we conduct ablation experiments on LLaVA-1.5-7B, with results reported

Table 5: Ablation experiments on the visual subspace construction, with the best results highlighted in **bold**.

Method	Acc $\uparrow$	F1 $\uparrow$
-	84.80	85.50
P_1	83.60	81.43
P_2	83.47	81.20
P_3	86.03	84.98
P_4	85.97	86.42
VRR	86.00	86.45

in Table 5. “-” denotes the baseline without VRR. P_1 directly adds the initial visual representation to the visual representations at all layers. P_2 applies the same direct addition only to the first half of the LLM layers. P_3 constructs a subspace from the visual representation of each corresponding layer. P_4 constructs a subspace from the full initial representation, incorporating both visual and textual representations. As shown in Table 5, P_1 and P_2 lead to performance degradation, suggesting that directly injecting the initial visual representation may disturb the original structure of layer-wise visual representations. P_3 achieves the highest Acc but results in a negative gain in F1, indicating less stable performance. P_4 improves the overall performance by constraining the correction within a subspace. In comparison, our method improves both Acc and F1, achieving the best F1 score of 86.45, which verifies the effectiveness of regularizing within the selected initial visual subspace.

Table 6: Ablation experiments on target visual representation, with the best results highlighted in **bold**.

Layer	Acc $\uparrow$	F1 $\uparrow$
-	84.80	85.50
28	84.80	85.50
20	84.77	85.48
6	85.10	85.76
3	85.30	85.89
VRR	86.00	86.45

Target Representation. VRR uses the initial visual representation before LLM processing as the target representation, since it preserves the most compact visual information. We validate this choice through ablation experiments, as shown in Table 6. Following the layer partition in Sec-

tion 5.4, we construct target subspaces from visual representations after the 3rd, 6th, 20th, and 28th layers, and regularize subsequent visual representations within these subspaces. The results show that earlier target representations consistently yield better performance, with the initial visual representation achieving the best results.

6 Conclusion

In this paper, we study hallucinations in MLLMs from the perspective of visual representation evolution. We show that visual representations progressively disperse across LLM layers, deviating from the initial visual subspace and aligning with the textual subspace. Motivated by this finding, we propose Visual Representation Regularization (VRR), a training-free method that preserves core initial visual information through subspace-aware regularization. Experiments on various benchmarks demonstrate that VRR consistently mitigates hallucinations across different MLLMs and further enhances existing inference-time mitigation methods. This work offers a new representation-level perspective for understanding and reducing hallucinations in MLLMs.

Limitation

SVD introduces additional computational overhead, while the projection spaces constructed by our method are relatively coarse and may not adequately capture the complex nonlinear structure of visual representations. In addition, performance gains are marginal or even negative on some metrics, and the diversity of evaluated model families remains limited. Future work will focus on developing more computationally efficient and fine-grained approaches for constructing projection spaces, analyzing the causes of negative performance changes, and evaluating the method across a broader range of model families.

Acknowledgment

This work was funded by NSFC (NO. 62206135, 62225604), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), the Guangdong Basic and Applied Basic Research Foundation (No.2026A1515011147) and the Fundamental Research Funds for the Central Universities (Nankai University, 070-63233085/63263251/63261032). Computation was supported by the Supercomputing Center of Nankai University.

References

- Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29915–29926.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. [Hallucination of multimodal large language models: A survey](#). *Preprint*, arXiv:2404.18930.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. [Halc: Object hallucination reduction via adaptive focal-contrast decoding](#). *Preprint*, arXiv:2403.00425.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pages 54158–54183.
- Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. 2024. [Explainable and interpretable multimodal large language models: A comprehensive survey](#). *Preprint*, arXiv:2412.02104.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2026. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385. IEEE.
- Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2025. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3488–3501.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.
- Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. 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)*, pages 25004–25014.
- Andrew V Knyazev and Merico E Argentati. 2002. Principal angles between subspaces in an a-based scalar product: algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2008–2040.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.
- Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. 2025. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4048–4080.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A. Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, Ying Shen, Barry Menglong Yao, Zhiyang Xu, Qin Liu, Yuxiang Zhang, Yan Sun, Shilong Liu, Li Shen, Hongxuan Li, and 2 others. 2025. [A survey on mechanistic interpretability for multi-modal foundation models](#). *Preprint*, arXiv:2502.17516.

- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. [A survey on hallucination in large vision-language models](#). *Preprint*, arXiv:2402.00253.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2024d. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024e. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. 2026. Revealing and enhancing core visual regions: Harnessing internal attention dynamics for hallucination mitigation in lvlms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 15214–15230.
- Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 57172–57189.
- Anirudh Phukan, Divyansh Divyansh, Harshit Kumar Morj, Vaishnavi Vaishnavi, Apoorv Saxena, and Koustava Goswami. 2025. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9661–9675.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. [A survey of multilingual large language models](#). *Patterns*, 6(1):101118.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pages 13088–13110.
- Enguang Wang, Qiang Wang, Yuanchen Wu, Ke Yan, Xinbin Yuan, Shouhong Ding, Xialei Liu, and Ming-Ming Cheng. 2026. [Predictive regularization against visual representation degradation in multimodal large language models](#). *Preprint*, arXiv:2603.20808.
- Kaishen Wang, Hengrui Gu, Meijun Gao, and Kaixiong Zhou. 2025. Damo: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 56764–56780.
- Dingkang Yang, Dongling Xiao, Jinjie Wei, Mingcheng Li, Zhaoyu Chen, Ke Li, and Lihua Zhang. 2025a. Improving factuality in large language models via decoding-time hallucinatory and truthful comparators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25606–25614.
- Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025b. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14635–14645.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051.
- Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. 2026. [Prefill-time intervention for mitigating hallucination in large vision-language models](#). *Preprint*, arXiv:2604.25642.
- Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024. [Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms](#). *Preprint*, arXiv:2411.09968.

- Li'an Zhong, Ziqiang He, Jibin Zheng, Jin Li, Z. Jane Wang, and Xiangui Kang. 2026. [Adaiat: Adaptively increasing attention to generated text to alleviate hallucinations in lvlm](#). *Preprint*, arXiv:2603.04908.
- Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025. *Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding*. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1615–1624.
- Xianwei Zhuang, Zhihong Zhu, Zhanpeng Chen, Yuxin Xie, Liming Liang, and Yuexian Zou. 2024. *Game on tree: Visual hallucination mitigation via coarse-to-fine view tree and game theory*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17984–18003.

A Proof of Eq. (8)

The expression of $\mathcal{J}(\widehat{H}^l)$ is known as Eq. (7). Taking the derivative of $\mathcal{J}(\widehat{H}^l)$ yields:

$$\frac{\partial \mathcal{J}(\widehat{H}^l)}{\partial \widehat{H}^l} = 2(\widehat{H}^l - H^l) + 2\frac{\alpha}{1-\alpha}(\widehat{H}^l - H)P \quad (9)$$

By setting $\frac{\partial \mathcal{J}(\widehat{H}^l)}{\partial \widehat{H}^l} = 0$, we obtain:

$$\widehat{H}^l = (H^l + \frac{\alpha}{1-\alpha}HP)(I + \frac{\alpha}{1-\alpha}P)^{-1} \quad (10)$$

By Eq. (6), V_k , as an orthonormal basis, satisfies:

$$I_k = (V_k)^T V_k \quad (11)$$

where I_k is the identity matrix of shape k .

According to the Woodbury matrix identity, for conformable matrices A, U, C , and V , we have:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}. \quad (12)$$

Let $A = I, U = V_k, C = \frac{\alpha}{1-\alpha}I_k, V = (V_k)^T$ and combine Eq. (6), Eq. (11) and Eq. (12), then:

$$(I + \frac{\alpha}{1-\alpha}P)^{-1} = I - \alpha P. \quad (13)$$

Substituting Eq. (13) back into Eq. (10) and using the property $P = P^2$ of the projection matrix yields:

$$\widehat{H}^l = H^l + \alpha(H - H^l)P, \quad (14)$$

which is equivalent to Eq. (8).

B Experiments Details

B.1 Benchmarks

CHAIR. CHAIR evaluates object hallucination by comparing generated captions with ground-truth data. It identifies objects mentioned in the captions that do not appear in the image and quantifies hallucination by computing their proportion. It’s metrics includes caption-level (CHAIR_S) and instance-level (CHAIR_I):

$$\text{CHAIR}_S = \frac{|\{\text{caption w/ hallucinated objects}\}|}{|\{\text{all captions}\}|}, \quad (15)$$

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \quad (16)$$

We randomly sample 500 images from the COCO 2014 validation set (Lin et al., 2014) and repeat the experiments 5 times using different random seeds. We employ the prompt “Please describe this image in detail” to query the MLLMs for image descriptions.

POPE. The POPE benchmark evaluates object hallucination by asking whether the model can correctly answer the presence of objects in an image. Each image is associated with questions such as “Is there a <object> in the image?”.

B.2 Experimental Setting Details

B.2.1 MLLMs.

All selected models consist of a vision encoder, a connector, and an LLM. They employ different strategies for generating vision tokens. LLaVA-1.5-7B directly encodes the original image into vision tokens. mPLUG-Owl2-7B introduces a visual abstractor to compress the vision tokens, thereby enhancing the understanding of global image semantics. In contrast, LLaVA-NeXT-7B and LLaVA-NeXT-13B jointly encodes the original image together with multiple image patches, generating a larger number of vision tokens to enhance the model’s fine-grained visual perception.

B.2.2 Baselines

VCD focuses on decoding refinement by comparing the output distributions between original and perturbed visual inputs, thereby reducing reliance on statistical bias and unimodal priors. VHR focuses on improving the attention module by enhancing the role of visually grounded attention heads to mitigate hallucinations. These methods represent two mainstream paradigms for hallucination mitigation. Nullu projects the input features into the null space spanned by hallucinated embedding features to reduce object hallucination. We incorporate VRR into the above baselines to demonstrate the superiority and robustness of our approach.

B.3 Results

B.3.1 Generation Quality Analysis

In addition to the original CHAIR_i and CHAIR_s , we report caption length, the number of mentioned objects, object precision (P), recall (R), and F1, as well as the standard captioning metrics SPICE and CLIPScore. Using mPLUG-Owl2-7B, the results are shown in Table 7. VRR substantially reduces CHAIR_i and CHAIR_s without sacrificing

Table 7: Generation quality analysis on mPLUG-Owl2-7B, with the best results highlighted in **bold**.

Method	CHAIR _i ↓	CHAIR _s ↓	Length	Mentioned objects	P↑	R↑	F1↑	SPICE↑	CLIPScore↑
Greedy	57.72	16.90	92.00	3.44	75.40	81.58	75.16	0.1838	80.25
w/ VRR	40.40	12.12	82.18	2.88	83.20	76.95	77.10	0.1929	80.45

generation quality. Despite generating shorter captions and mentioning fewer objects, VRR achieves significant improvements in precision and F1, indicating better visual grounding. The marginal improvements in SPICE and CLIPScore further suggest that these gains are not simply attributable to more conservative generation. We qualitatively demonstrate the hallucination mitigation capability of VRR, as shown in Figure 5. VRR suppresses the generation of hallucinated terms while preserving the coherence of the original semantics.

B.3.2 Generalization Experiment

Table 8: Comparison of different baselines across benchmarks using LLaVA-NeXT-13B. The best results are highlighted in **bold**.

Baselines	MM	Hallusion	MMHal
Greedy	75.72	51.11	2.031
w/ VRR	77.73	53.23	2.073
VHR	74.29	50.48	2.333
w/ VRR	74.24	50.57	2.396

We additionally conduct experiments on LLaVA-NeXT-13B using both Greedy and VHR. We report MMBench overall accuracy (MM) (Liu et al., 2024e), HallusionBench all accuracy (Hallusion) (Guan et al., 2024), and MMHal-Bench average score (MMHal) (Sun et al., 2024). The results are shown in Table 8. VRR consistently achieves competitive performance under both baselines, with particularly significant improvements under Greedy. It demonstrates that VRR is also effective on broader multimodal hallucination benchmarks, further validating its generalizability.

B.3.3 Comparative Experiment

We compare VRR directly with PTI (Zhang et al., 2026) on the MSCOCO subset of the POPE benchmark using LLaVA-1.5-7B. PTI intervenes in the prefill stage to mitigate hallucinations. It is also compatible with other hallucination mitigation methods and features plug-and-play functionality. We then evaluate the compatibility of both methods. In addition to Greedy and Nullu used in the main text, we introduce two additional hallucination mitigation methods, OPERA (Huang et al.,

Table 9: Comparison of different baselines with PTI and VRR on the MSCOCO subset of the POPE benchmark using LLaVA-1.5-7B.

Method	Acc ↑	F1 ↑
Greedy	83.72	84.77
w/ PTI	84.09	85.02
w/ VRR	84.79	85.58
Nullu	84.80	85.31
w/ PTI	84.96	85.34
w/ VRR	85.69	85.88
OPERA	83.80	84.83
w/ PTI	84.12	85.04
w/ VRR	84.87	85.64
PAI	84.84	85.41
w/ PTI	84.69	85.19
w/ VRR	85.61	85.85

2024) and PAI (Liu et al., 2024d). The performance improvements over each baseline are shown in Table 9. The results show that VRR consistently outperforms PTI across all baselines. In particular, when combined with PAI, PTI even degrades the performance compared with using PAI alone, whereas VRR consistently improves upon the original PAI baseline, further demonstrating its strong plug-and-play compatibility. Moreover, when used alone, VRR achieves a higher F1 score than existing training-free methods such as Nullu and PAI, further demonstrating its superior effectiveness.

B.4 Further analysis

Table 10: Ablation experiments of γ conducted on the three models, with the best results highlighted in **bold**.

γ	mPLUG-Owl2-7B		LLaVA-1.5-7B		LLaVA-NeXT-7B	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
-	77.40	80.43	84.80	85.50	88.97	88.70
0.1	82.83	83.85	85.80	86.32	89.53	89.16
0.2	81.97	83.31	85.67	86.19	89.60	89.22
0.5	83.13	84.19	86.00	86.45	89.13	88.64
1.0	81.33	82.99	86.07	86.38	88.77	88.06

Ablation experiments of γ . Table 10 presents the ablation results of γ , which denotes the proportion of principal directions retained from all non-zero principal components. The setting “-” indicates the baseline without VRR. The optimal value

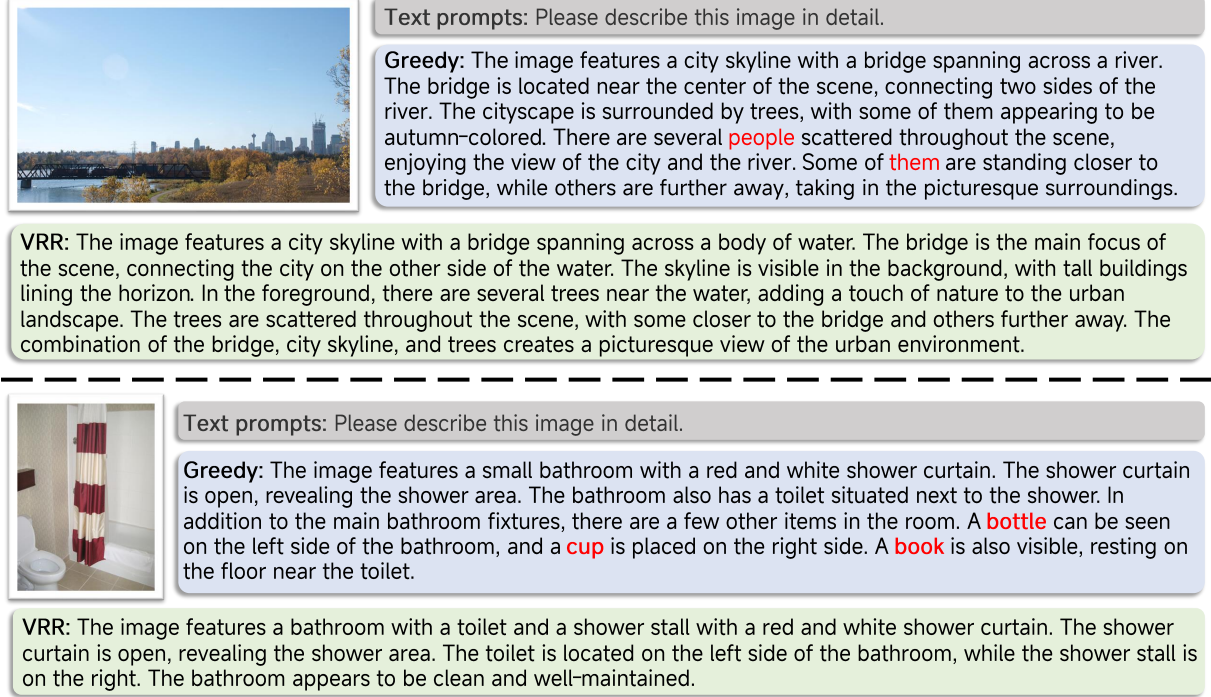


Figure 5: Examples of VRR mitigating hallucinated content using mPLUG-Owl2-7B. Red words indicate hallucinated terms.

of γ varies across models, and mPLUG-Owl2-7B achieves larger gains than the other two models, attaining the best performance at $\gamma = 0.5$. For LLaVA-1.5-7B, Acc is highest at $\gamma = 1.0$, while F1 is highest at $\gamma = 0.5$. LLaVA-NeXT-7B achieves the best performance at $\gamma = 0.2$. When all non-zero principal components are retained ($\gamma = 1.0$), the performance of LLaVA-NeXT-7B decreases, suggesting that retaining all non-zero principal components may introduce noise. To preserve the dominant information while retaining as little noise as possible, we finally set γ to 0.5, 0.5, and 0.2 for mPLUG-Owl2-7B, LLaVA-1.5-7B, and LLaVA-NeXT-7B, respectively.

Computational Efficiency. We analyze the computational efficiency of VRR on the POPE-Popular benchmark using LLaVA-1.5-7B, and report the latency of the first forward pass (Latency), peak GPU memory consumption (Memory), and the corresponding accuracy (Acc).

Table 11: The relationship between layer count and computational efficiency on LLaVA-1.5-7B.

Layer	-	5	19	32
Latency (s)	0.158	+0.071	+0.076	+0.082
Memory (GB)	13.83	+0.08	+0.08	+0.08
Acc (%)	84.80	+0.30	+1.20	+1.20

We first analyze the layer-wise performance-efficiency trade-off. The results are shown in Table 11. Performance becomes stable after the 19th layer, while latency and memory consumption remain nearly unchanged, indicating that layer-wise regularization has a negligible impact on efficiency.

Table 12: The relationship between γ and latency on LLaVA-1.5-7B.

γ	Latency (s)	Acc (%)
-	0.158	84.80
0.1	+0.021	+1.00
0.5	+0.057	+1.20
1.0	+0.122	+1.27

VRR constructs the projection matrix via Partial SVD, with γ controlling the number of retained basis vectors and the resulting latency. We further analyze the relationship between γ and latency on LLaVA-1.5-7B, with the results shown in Table 12. As γ increases, the latency gradually increases, while the accuracy remains nearly unchanged once $\gamma \geq 0.5$. This suggests that retaining more basis vectors beyond this point provides limited accuracy gains at the cost of increased latency.