Mitigating Hallucinations in Large Vision-Language Models without Performance Degradation

Xingyu Zhu[1,2], Junfeng Fang[2*], Shuo Wang[1],
Beier Zhu[1], Zhicai Wang[1], Yonghui Yang[2], Xiangnan He[1*]

[1] MoE Key Lab of BIPC, University of Science and Technology of China,
[2] National University of Singapore

Abstract

Large Vision-Language Models (LVLMs) exhibit powerful generative capabilities but frequently produce hallucinations that compromise output reliability. Fine-tuning on annotated data devoid of hallucinations offers the most direct solution, while its high computational cost motivates recent representation-based methods, which focus on mitigating hallucinatory components within hidden representations. Though efficient, we empirically observe that these methods degrade general generation capacity due to incomplete extraction of hallucination components and non-selective parameter updates. To address these limitations, we propose MPD, a dual-stage framework for mitigating hallucinations without performance degradation. Specifically, our MPD relies on two essential factors: (1) semantic-aware component disentanglement to extract pure hallucination components, and (2) interpretable parameter updates that selectively modify parameters most relevant to hallucination. Extensive experiments demonstrate that MPD achieves state-of-the-art performance, reducing hallucinations by 23.4% while maintaining 97.4% of general generative capability as evaluated on LLaVA-Bench and MME, with no additional computational cost.

1 Introduction

Recent advances in large vision-language models (LVLMs) (Zhu et al., 2024; Liu et al., 2023b; Ye et al., 2023; Dai et al., 2023; Bai et al., 2023; Zhu et al., 2026c) have demonstrated remarkable capabilities in cross-modal understanding and generation. However, these models exhibit a persistent limitation known as hallucination phenomena (Gunjal et al., 2024; Liu et al., 2024a; Zhu et al., 2026a), *i.e.*, a critical divergence where generated textual descriptions systematically misrepresent visual content. Typical manifestations include fabricating non-existent objects, misattributing visual properties, or hallucinating erroneous spatial relationships within images. These systematic errors not only undermine practical applications requiring precise vision-language alignment, but also pose significant risks for misinformation propagation and safety-critical deployment scenarios (Liu et al., 2024b; Chen et al., 2024b).

Conventional approaches (Chen et al., 2024c; Leng et al., 2024; Chen et al., 2024a) to hallucination mitigation primarily rely on labor-intensive dataset curation through manual annotation to filter hallucinatory content for fine-tuning, inherently suffering from time-consuming bottlenecks. Hence, recent research has shifted toward representation intervention (Yang et al., 2025; Uppaal et al., 2024; Liu et al., 2024c), introducing a dual-stage paradigm. The first step focuses on extracting hallucinatory components from hidden representations by contrasting the model's internal activations when it produces hallucinated versus faithful outputs. Building on this, the subsequent step updates model parameters to minimize the influence of the extracted hallucinatory components. This paradigm eliminates the dependency on exhaustive dataset reconstruction while enabling real-time hallucination mitigation.

While effective, current representation intervention methods (Yang et al., 2025; Li et al., 2024; Turner et al., 2023) suffer from a critical limitation: post-intervened LVLMs would lose general generative capabilities. As quantified in Figure 1 (a), post-intervened LVLMs exhibit more semantic incoherence and lexical repetition rates compared to original models. We attribute this limitation to two flaws in current paradigms: First, during the extraction of hallucinatory components, these components are typically highly coupled with general semantic components. Existing methods overlook such shared components, leading to the unintended removal of general semantics. As shown in the

*Corresponding authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1995–2009

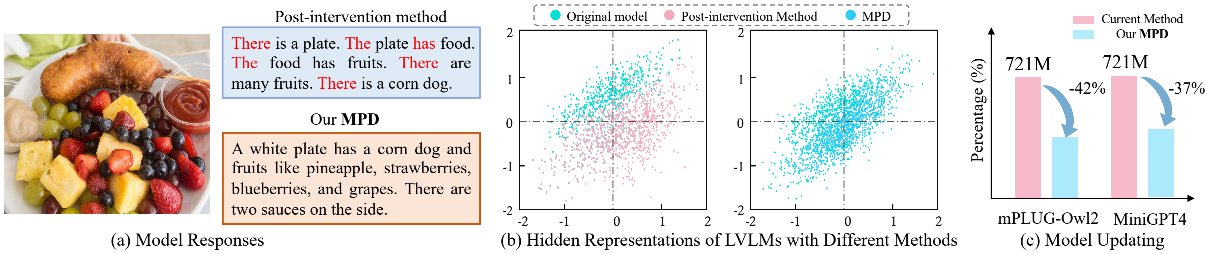

Figure 1: **Comparison between conventional representation-intervention methods and MPD. (a)** Model responses to the same input image show that traditional methods produce repetitive and incoherent descriptions, while MPD generates coherent and informative outputs. **(b)** Post-intervention hidden representations from prior methods(Yang et al., 2025; Liu et al., 2024c) drift away from the original distribution, **(c)** MPD aligns the hidden representations more closely with the original distribution. **(d)** MPD also requires significantly fewer parameter updates and FLOPs, enabling more efficient hallucination mitigation.

left part of Figure 1 (b), the distribution of representations in post-intervened LVLMs diverges significantly from that of the original model. Second, during parameter updating, substantial perturbations are often applied to all weights within the targeted layers, resulting in the modification of hundreds of millions of parameters, as shown in Figure 1 (c). Such large-scale updates inevitably induce overfitting and severely disrupt the original parameter distributions.

To address these issues, we propose **MPD**, a *dual-stage framework* that mitigates hallucinations without degrading generative performance. In the first stage, we construct contrastive query pairs using an auxiliary large language model (LLM), where each pair comprises a hallucination-inducing question and a semantically equivalent non-hallucinatory question. By analyzing the model's representations for both cases, we disentangle hallucination components through orthogonal projection. This design effectively prevents contamination between hallucination and general semantics, as evidenced by the well-aligned post-intervention distributions in Figure 1(b); theoretical analysis is provided in Section 3.4. In the second stage, we identify parameters most correlated with hallucination components via cosine similarity and selectively update only these critical parameters. As shown in Figure 1(c), this selective update substantially reduces parameter perturbations by 42% on mPLUG-Owl2 (Ye et al., 2023) and 37% on MiniGPT4 (Zhu et al., 2024), thus preserving the model's general capability. We term this approach **MPD**, highlighting its ability to mitigate hallucinations without performance degradation.

To validate MPD's efficacy, we conduct extensive experiments on five datasets (*i.e.*, CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023), MME (Liu et al., 2023a), LLaVA-Bench (Liu et al., 2023a)) across three prevailing base LVLMs. Quantitative results demonstrate that MPD achieves advancing hallucination mitigation while preserving 97.4% of LLMs' generative capabilities. Notably, MPD introduces no extra computational cost during inference compared to the original model. This breakthrough bridges the efficiency-effectiveness gap in hallucination mitigation, offering practical scalability for real-world LVLM deployment.

2 Related Works

2.1 Hallucination in LVLMs

LVLMs integrate powerful LLMs with visual backbones via modality alignment modules (Liu et al., 2023b; Zhu et al., 2024; Chen et al., 2023; Zhu et al., 2026b; Zhu et al.), have demonstrated remarkable performance across various multimodal tasks (He et al., 2024; Bai et al., 2023). However, these models are prone to object hallucination (Bai et al., 2024), where the generated response includes non-existent objects or incorrect attributes inconsistent with the visual input (Rohrbach et al., 2018; Li et al., 2023). To mitigate OH, several approaches rely on fine-tuning with additional supervision, including reinforcement learning from human feedback (Sun et al., 2024; Gunjal et al., 2024), auxiliary annotation (Jiang et al., 2024), or the inclusion of noisy/negative samples (Liu et al., 2024a). Although effective, such methods require high computational costs and hinder scalability. To reduce the overhead of fine-tuning, recent studies have explored training-free alternatives. These include decoding-time constraints (Leng et al., 2024; Zhang et al., 2025) or post-hoc revision modules (Yin et al., 2023; Chen et al., 2024c) that detect and correct hallucinated tokens using external visual grounding. However, these methods either introduce extra inference latency or depend on auxiliary models. In contrast, our method adopts a training-free, editing-based approach: it identifies hallucination-inducing directions in hidden representations and directly edits the model's internal parameters to suppress them. This enables efficient and consistent hallucination reduction.

2.2 Subspace Projection and Model Editing

Subspace projection has emerged as a powerful parameter-editing strategy for steering model behavior, enabling the suppression of undesired semantics (*e.g.*, toxicity, bias, hallucination) by identifying and removing specific directions in the representation or weight space (Turner et al., 2023; Fang et al., 2025). In the multimodal context, Nullu (Yang et al., 2025) applies this idea to LVLMs by projecting weights onto the null space of hallucination-related directions estimated from hidden features, achieving hallucination suppression via global weight editing. More generally, parameter-modifying model editing techniques adjust internal weights either through meta-learning (*e.g.*, MEND (Mitchell et al., 2022), InstructEdit (Zhang et al., 2024)) or locate-and-edit strategies (*e.g.*, ROME (Meng et al., 2022), AnyEdit (Jiang et al., 2025)), typically relying on key-value memory, gradient updates, or causal tracing to identify editable components. Recent extensions to multimodal models include works such as OPERA (Huang et al., 2024) and VCD (Leng et al., 2024), which incorporate vision-aware attention refinement or mask-guided cross-modal grounding, often acting at the decoding or cross-attention layers. In contrast, our method introduces a fine-grained orthogonal subspace projection mechanism guided by contrastive pairs between faithful and hallucinated hidden representations. Instead of modifying all parameters, we edit only those weights most aligned with hallucination-specific directions, achieving effective mitigation with minimal disruption to the model's generation behavior.

3 Method

We first outline hallucination behavior in LVLMs (Section 3.1), then extract hallucinated components via hidden state differencing (Section 3.2). Next, we introduce selective parameter editing to suppress hallucination by projecting key weights away from the hallucination subspace (Section 3.3), and provide theoretical justification via orthogonal projection onto the faithful subspace (Section 3.4).

3.1 Preliminary

A typical LVLM (Zhu et al., 2024; Chen et al., 2023; Liu et al., 2023b; Ye et al., 2023) consists of a vision encoder, a text encoder, and a large language model as the decoder. Given an image and a textual input (*e.g.*, a question), the vision encoder extracts visual features, which are projected into the same embedding space as text and concatenated for auto-regressive decoding. However, LVLMs often hallucinate (Yin et al., 2023; Leng et al., 2024), mentioning objects not present in the image. To address this, we build contrastive vision-language pairs sharing the same image but differing in text: one prompt (x_i^-) includes hallucinated content, while the other (x_i^+) provides a faithful visual description grounded in the image. The full dataset is denoted as $\mathcal{D} = \{(x_i^+, x_i^-)\}_{i=1}^N$, and the details are provided in Appendix B.

3.2 Hallucination Component Extraction

Extracting hallucinatory components is the first step of our method. Since hallucinations arise from subtle deviations in hidden representations between faithful and hallucinated responses, we analyze token-level features across transformer layers $\ell \in \{L_0, \dots, L\}$ in the base LLM of the LVLM. For each image-caption pair, we obtain token embeddings from both faithful and hallucinated descriptions, denoted as $\mathbf{x}_{i,\ell}^+$ and $\mathbf{x}_{i,\ell}^-$, respectively. To obtain a compact representation, we average token embeddings across the sequence. The resulting vectors are stacked into matrices $\mathbf{X}_\ell^+, \mathbf{X}_\ell^- \in \mathbb{R}^{N \times D}$, where D is the feature dimension. We then model the hallucinated representation as a composition of grounded semantics, hallucinatory component, and residual noise:

$$\mathbf{X}_\ell^- = \mathbf{X}_\ell^{\text{real}} + \mathbf{X}_\ell^{\text{hall}} + \boldsymbol{\epsilon}^-, \tag{1}$$

where $\mathbf{X}_\ell^{\text{real}}$ aligns with the image, $\mathbf{X}_\ell^{\text{hall}}$ captures hallucinated semantics, and $\boldsymbol{\epsilon}^-$ is residual noise. This decomposition forms the basis for hallucinatory component extraction.

Since $\mathbf{X}_\ell^+$ encodes faithful descriptions, its row space approximates the grounded semantic subspace. We perform singular value decomposition (SVD) on $\mathbf{X}_\ell^+$:

$$\mathbf{X}_\ell^+ = \mathbf{U}_\ell \boldsymbol{\Sigma}_\ell \mathbf{V}_\ell^\top, \tag{2}$$

where $\mathbf{U}_\ell \in \mathbb{R}^{N\times C}$ contains the top-C left singular vectors. The projection matrix onto the grounded subspace is given by $\mathbf{P}_\ell = \mathbf{U}_\ell \mathbf{U}_\ell^\top \in \mathbb{R}^{N\times N}$. We apply $\mathbf{P}_\ell$ to $\mathbf{X}_\ell^-$ to obtain the grounded component $\hat{\mathbf{X}}_\ell^- = \mathbf{P}_\ell \mathbf{X}_\ell^-$ and the hallucinatory component is extract as:

$$\tilde{\mathbf{X}}_\ell = \mathbf{X}_\ell^- - \hat{\mathbf{X}}_\ell^- = (\mathbf{I} - \mathbf{P}_\ell)\mathbf{X}_\ell^-, \qquad (3)$$

where $\mathbf{I} \in \mathbb{R}^{D\times D}$ is the identity matrix. The hallucinatory component $\tilde{\mathbf{X}}_\ell$ lies orthogonal to the faithful subspace. The justification analysis are provided in Section 3.4.

Algorithm 1 Pipeline of **MPD**

Input: Paired data $\mathcal{D}$, target layers $\{\ell\}$ in LVLM.
1: **for** ℓ in $\{\ell\}$ **do**
2: Calculating hidden states of hallucinatory component $\mathbf{X}^-$ and faithful components $\mathbf{X}^+$.
3: Extracting the orthogonal components of the faithful subspace $\tilde{\mathbf{X}}_\ell$ via Eq. (3).
4: Selecting weight vectors $\{\mathbf{w}_\ell^{(i)}\}$ with top-K similarity with $\tilde{\mathbf{X}}_\ell$ via Eq. (4).
5: Calculating the projection matrix $\tilde{\mathbf{Q}}_\ell$ via Eq. (5).
6: Updating the model parameters via $\{\mathbf{w}_\ell^{(i)}\}$ via Eq. (6)
7: **end for**

3.3 Parameters Updating

In Section 3.2, we described extracting hallucinatory components from hidden states. We update the parameters most responsible for hallucinations.

Parameter selection. Given the hallucination component $\tilde{\mathbf{X}}_\ell = [\tilde{\mathbf{x}}_{\ell,1}, \ldots, \tilde{\mathbf{x}}_{\ell,N}]^\top \in \mathbb{R}^{N\times D}$ extracted at layer ℓ, we aim to identify the weight parameters that are most responsible for hallucinated generation. Let $\mathbf{W}_\ell \in \mathbb{R}^{L\times D}$ denote the weight matrix at layer ℓ, where each row $\mathbf{w}_\ell^{(i)}$ is a weight vector associated with a neuron. To quantify the alignment between each weight and hallucinated semantics, we compute the average cosine similarity between $\mathbf{w}_\ell^{(i)}$ and the $\tilde{\mathbf{x}}_{\ell,j}$:

$$s_i := \frac{1}{N}\sum_{j=1}^{N} \frac{\cos(\mathbf{w}_\ell^{(i)}, \tilde{\mathbf{x}}_{\ell,j})}{\|\mathbf{w}_\ell^{(i)}\| \cdot \|\tilde{\mathbf{x}}_{\ell,j}\|}. \qquad (4)$$

We then select the top-K weight vectors with the highest similarity scores s_i, and denote their indices by $\mathcal{I}_\ell^{\text{hall}} \subset \{1,\ldots,L\}$. These weights are considered most aligned with hallucinated directions and are the targets of our editing step.

Parameter editing. To suppress hallucination-related influence, we construct a projection operator that eliminates components lying in the hallucination subspace. Following (Fang et al., 2025), we define the projection matrix onto the orthogonal complement (null space) of $\tilde{\mathbf{X}}_\ell$ as:

$$\tilde{\mathbf{Q}}_\ell = \mathbf{I} - \tilde{\mathbf{X}}_\ell^\top (\tilde{\mathbf{X}}_\ell \tilde{\mathbf{X}}_\ell^\top)^{-1} \tilde{\mathbf{X}}_\ell. \qquad (5)$$

This projection operator removes directions spanned by $\tilde{\mathbf{X}}_\ell$. By projecting onto this complement, we can remove the semantic directions associated with hallucinatory component, yielding purified features that better reflect grounded visual evidence. We apply $\tilde{\mathbf{Q}}_\ell$ to selectively edited weights. For each $i \in \mathcal{I}_\ell^{\text{hall}}$, we update the model parameters as follows:

$$\mathbf{w}_\ell^{(i)} \leftarrow \tilde{\mathbf{Q}}_\ell \mathbf{w}_\ell^{(i)}. \qquad (6)$$

This targeted projection suppresses hallucination-inducing capacity with minimal impact on unrelated behaviors, as only a small subset of weights is modified. The complete procedure of our **MPD** is summarized in Algorithm 1.

3.4 Theoretical Analysis

Sections 3.2 and 3.3 have outlined how to extract hallucination features and suppress them via selective weight projection. We now provide a theoretical justification for using the projection-based residual as a principled estimator of hallucination-specific components.

Justification of projection-based residual for hallucination extraction. To extract the hallucinatory components, we decompose $\mathbf{X}_\ell^{\text{hall}}$ into a parallel part $\mathbf{X}_\ell^{\text{hall},\parallel} = \mathbf{P}_\ell \mathbf{X}_\ell^{\text{hall}}$ within the subspace of faithful representations $\mathbf{X}_\ell^+$, and an orthogonal part $\mathbf{X}_\ell^{\text{hall},\perp} = (\mathbf{I} - \mathbf{P}_\ell)\mathbf{X}_\ell^{\text{hall}}$, corresponding to disentangled hallucinatory components. We define $\tilde{\mathbf{X}}_\ell = (\mathbf{I} - \mathbf{P}_\ell)\mathbf{X}_\ell^-$ as the extracted hallucinatory components from the response representation $\mathbf{X}_\ell^-$. As a comparison, we consider the naive difference-based estimate $\tilde{\mathbf{X}}_\ell^{\text{diff}} = \mathbf{X}_\ell^- - \mathbf{X}_\ell^+$, which entangles hallucination with shared semantics.

Proposition 1. *For any $\mathbf{X}_\ell^{hall,\parallel} = \mathbf{P}_\ell \mathbf{X}_\ell^{hall}$, the extracted hallucinatory components $\tilde{\mathbf{X}}_\ell = (\mathbf{I} - \mathbf{P}_\ell)\mathbf{X}_\ell^-$ provide a more accurate estimation of the disentangled hallucinatory signal $\mathbf{X}_\ell^{hall,\perp}$ than the naive difference-based estimate $\tilde{\mathbf{X}}_\ell^{diff} = \mathbf{X}_\ell^- - \mathbf{X}_\ell^+$, in terms of expected squared error:*

$$\mathbb{E}\|\tilde{\mathbf{X}}_\ell - \mathbf{X}_\ell^{hall,\perp}\|_F^2 \le \mathbb{E}\|\tilde{\mathbf{X}}_\ell^{diff} - \mathbf{X}_\ell^{hall,\perp}\|_F^2.$$

Proof. The extracted hallucinatory components is $\tilde{\mathbf{X}}_\ell = (\mathbf{I} - \mathbf{P}_\ell)\mathbf{X}_\ell^-$, where $\mathbf{P}_\ell$ projects onto the faithful subspace spanned by $\mathbf{X}_\ell^+$. For $\mathbf{X}_\ell^- = \mathbf{X}_\ell^{\text{real}} + \mathbf{X}_\ell^{\text{hall}} + \boldsymbol{\epsilon}^-$, with assuming $\boldsymbol{\epsilon}^- \sim \mathcal{N}(0, \sigma_-^2\mathbf{I})$, and $\mathbf{X}_\ell^{\text{real}}$ lies in the faithful subspace ($\mathbf{P}_\ell\mathbf{X}_\ell^{\text{real}} \approx \mathbf{X}_\ell^{\text{real}}$), the error is:

$$\tilde{\mathbf{X}}_\ell - \mathbf{X}_\ell^{\text{hall},\perp} \approx (\mathbf{I} - \mathbf{P}_\ell)\boldsymbol{\epsilon}^-,$$

yielding:

$$\mathbb{E}\|\tilde{\mathbf{X}}_\ell - \mathbf{X}_\ell^{\text{hall},\perp}\|_F^2 = \sigma_-^2(D - C)N.$$

The difference-based residual is $\tilde{\mathbf{X}}_\ell^{\text{diff}} = \mathbf{X}_\ell^- - \mathbf{X}_\ell^+$, with $\mathbf{X}_\ell^+ = \mathbf{X}_\ell^{\text{real}} + \boldsymbol{\epsilon}^+$, $\boldsymbol{\epsilon}^+ \sim \mathcal{N}(0, \sigma_+^2\mathbf{I})$. The error is:

$$\tilde{\mathbf{X}}_\ell^{\text{diff}} - \mathbf{X}_\ell^{\text{hall},\perp} = \mathbf{X}_\ell^{\text{hall},\parallel} + \boldsymbol{\epsilon}^- - \boldsymbol{\epsilon}^+.$$

Based on these, we have:

$$\mathbb{E}\|\tilde{\mathbf{X}}_\ell^{\text{diff}} - \mathbf{X}_\ell^{\text{hall},\perp}\|_F^2 = \|\mathbf{X}_\ell^{\text{hall},\parallel}\|_F^2 + \sigma_-^2 DN + \sigma_+^2 DN.$$

Since $\sigma_-^2(D - C)N < \sigma_-^2 DN$, and $\|\mathbf{X}_\ell^{\text{hall},\parallel}\|_F^2 \geq 0$, $\sigma_+^2 \geq 0$ contribute additional non-negative terms, we have:

$$\sigma_-^2(D - C)N < \|\mathbf{X}_\ell^{\text{hall},\parallel}\|_F^2 + \sigma_-^2 DN + \sigma_+^2 DN.$$

Thus, $\tilde{\mathbf{X}}_\ell$ provides a more accurate estimation. See Appendix A for details. □

4 Experiments

In this section, we conduct experiments to address the following research questions (RQ):

- **RQ1:** Does MPD effectively mitigate object hallucinations in LVLMs across multiple settings, including sentence-level, instance-level, and representation-level metrics?
- **RQ2:** Does our method maintain the perception and reasoning abilities of LVLMs while suppressing hallucinations?
- **RQ3**: How do the edited layer ranges, sample sizes, retained SVD components, and selected weights affect the hallucination suppression?

4.1 Experimental Setup

We begin by outlining the evaluation benchmarks, metrics, and baseline methods used to assess our approach across multiple LVLMs.

Base models and baselines. Our experiments are conducted on three representative LVLMs: **MiniGPT-4 V2** (Zhu et al., 2024), **LLaVA-1.5-7B** (Liu et al., 2023b), and **mPLUG-Owl2** (Ye et al., 2023). For comparison, we include several state-of-the-art object hallucination (OH) mitigation methods: DoLa (Chuang et al., 2023), OPERA (Huang et al., 2024), VCD (Leng et al., 2024), LURE (Zhou et al., 2023), HALC (Chen et al., 2024c), and Nullu (Yang et al., 2025). We use LURE (Zhou et al., 2023) as the paired data following the previous work (Yang et al., 2025), and prompt all methods with "*Please describe this image in detail.*".

Evaluation benchmarks and metrics. We evaluate OH mitigation performance using the following three categories of benchmarks. **MSCOCO-based metrics.** We adopt CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023) on the MSCOCO (Lin et al., 2014) dataset to evaluate hallucination in image descriptions. CHAIR includes sentence-level (CHAIR_S) and instance-level (CHAIR_I) hallucination metrics, with lower scores indicating better grounding. POPE measures LVLMs' ability to correctly answer yes/no object presence queries under different negative sampling strategies. **MME.** The Multimodal Large Language Model Evaluation (MME) benchmark (Fu et al., 2023) is used to assess perception and reasoning abilities across 14 tasks. We follow prior work (Yin et al., 2023; Huang et al., 2024; Chen et al., 2024c) and focus on four hallucination-relevant subsets: "Existence", "Count", "Position", and "Color". **LLaVA-Bench.** The LLaVA-Bench (Liu et al., 2023a) benchmark comprises human-curated images and a diverse set of questions to evaluate open-ended captioning, reasoning, and conversational abilities. Details are provided in Appendix B.

4.2 Performance on Object Hallucinations (RQ1)

To evaluate MPD's effectiveness in mitigating object hallucinations, we conduct experiments on CHAIR and POPE, followed by an analysis of hidden representation alignment.

Results on CHAIR. We evaluate the effectiveness of our method on object hallucination using the CHAIR benchmark, which includes both sentence-level (CHAIR_S) and instance-level (CHAIR_I) hallucination metrics, as well as BLEU for caption quality. Table 1 summarizes the performance across three representative LVLMs.

- **Obs 1: MPD significantly reduces sentence-

Table 1: CHAIR evaluation results on MSCOCO dataset of LVLMs (MiniGPT-4, mPLUG-Owl2, and LLaVA-1.5-7B) with different methods for mitigating OH. Lower CHAIR_S and CHAIR_I indicate less OH. Higher BLEU generally represents higher captioning quality. We use 64 as the max token number in this experiment. Bold indicates the best result of all methods.

Method	MiniGPT-4			mPLUG-Owl2			LLaVA-1.5-7B		
	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	BLEU↑	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	BLEU↑	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$	BLEU↑
Greedy	32.40$_{\pm 2.20}$	12.20$_{\pm 0.42}$	14.57$_{\pm 0.11}$	22.90$_{\pm 0.90}$	8.62$_{\pm 0.11}$	15.01$_{\pm 0.24}$	20.40$_{\pm 2.80}$	7.08$_{\pm 0.33}$	15.72$_{\pm 0.10}$
DoLa	31.90$_{\pm 3.30}$	12.15$_{\pm 0.89}$	14.54$_{\pm 0.12}$	22.40$_{\pm 1.80}$	8.36$_{\pm 0.04}$	15.13$_{\pm 0.21}$	20.20$_{\pm 2.80}$	6.75$_{\pm 0.54}$	15.68$_{\pm 0.10}$
OPERA	29.70$_{\pm 0.30}$	11.96$_{\pm 0.29}$	14.82$_{\pm 0.05}$	20.07$_{\pm 2.07}$	7.18$_{\pm 0.39}$	15.41$_{\pm 0.12}$	17.50$_{\pm 0.50}$	6.07$_{\pm 0.32}$	16.02$_{\pm 0.02}$
VCD	29.00$_{\pm 2.80}$	12.64$_{\pm 1.19}$	14.42$_{\pm 0.01}$	22.80$_{\pm 0.80}$	8.68$_{\pm 0.17}$	15.14$_{\pm 0.13}$	20.30$_{\pm 1.10}$	7.28$_{\pm 0.10}$	14.53$_{\pm 0.01}$
LURE	27.88$_{\pm 2.25}$	10.20$_{\pm 0.85}$	15.03$_{\pm 0.11}$	21.27$_{\pm 0.06}$	7.67$_{\pm 0.16}$	15.65$_{\pm 0.15}$	19.48$_{\pm 2.35}$	6.50$_{\pm 0.38}$	15.97$_{\pm 0.01}$
HALC	25.20$_{\pm 2.00}$	9.42$_{\pm 0.41}$	14.91$_{\pm 0.13}$	18.80$_{\pm 1.20}$	7.00$_{\pm 0.01}$	15.33$_{\pm 0.24}$	16.90$_{\pm 2.10}$	5.72$_{\pm 0.55}$	16.02$_{\pm 0.04}$
Nullu	21.40$_{\pm 1.20}$	8.99$_{\pm 0.56}$	14.81$_{\pm 0.06}$	15.60$_{\pm 1.50}$	5.77$_{\pm 0.011}$	15.45$_{\pm 0.13}$	15.20$_{\pm 0.50}$	5.30$_{\pm 0.13}$	15.69$_{\pm 0.07}$
MPD	**19.40**$_{\pm 1.00}$	**7.50**$_{\pm 0.36}$	**14.98**$_{\pm 0.06}$	**14.00**$_{\pm 1.20}$	**4.99**$_{\pm 0.01}$	**16.06**$_{\pm 0.01}$	**12.80**$_{\pm 0.60}$	**4.20**$_{\pm 0.03}$	15.31$_{\pm 0.04}$

Table 2: Results on POPE. "Original" denotes direct inference using the original LVLMs, while "Nullu" refers to models enhanced with our proposed editing method.

Setting	Model	Method	Accuracy	Precision	Recall	F1 Score
random	MiniGPT4	Original	57.33	53.66	97.13	70.04
		Nullu	58.00	54.41	98.53	70.14
		MPD	**59.39**	**55.67**	**98.86**	**71.81**
	mPLUG-Owl2	Original	81.83	77.80	89.07	83.06
		Nullu	83.33	79.10	90.60	84.46
		MPD	**85.63**	**82.30**	**91.56**	**85.76**
	LLaVA-1.5-7B	Original	87.98	88.55	80.43	86.03
		Nullu	88.23	91.31	81.86	87.43
		MPD	**89.31**	**92.41**	**84.10**	**88.20**
popular	MiniGPT4	Original	56.63	53.66	97.13	69.13
		Nullu	52.66	51.38	98.53	67.57
		MPD	**53.96**	**52.24**	**98.75**	**68.53**
	mPLUG-Owl2	Original	75.77	70.35	89.07	78.61
		Nullu	77.47	71.75	90.60	80.08
		MPD	**79.57**	**74.79**	**91.80**	**81.58**
	LLaVA-1.5-7B	Original	84.68	81.51	81.43	83.32
		Nullu	85.35	84.35	82.13	84.93
		MPD	**86.37**	**85.82**	**83.93**	**85.52**
adversarial	MiniGPT4	Original	50.17	50.21	97.13	67.02
		Nullu	51.90	50.98	97.86	67.04
		MPD	**52.83**	**51.50**	**98.53**	**68.22**
	mPLUG-Owl2	Original	72.77	67.17	89.07	76.58
		Nullu	73.23	67.25	90.40	77.15
		MPD	**75.23**	**69.15**	**91.60**	**78.72**
	LLaVA-1.5-7B	Original	77.97	72.79	79.43	78.24
		Nullu	78.61	75.51	79.86	79.07
		MPD	**79.56**	**75.98**	**82.26**	**80.18**

level hallucinations while preserving caption quality. On all models, our method achieves the lowest CHAIR_S scores: 12.80, 19.40, and 14.00, outperforming strong baselines such as Nullu (15.20, 20.30, 15.30). BLEU scores remain competitive, with the highest score on mPLUG-Owl2 (16.06) and solid results on the other two, indicating reduced hallucination without sacrificing generation quality.

- **Obs 2: There is improvement with MPD in fine-grained hallucination mitigation.** In terms of CHAIR_I, our method outperforms DoLa, HALC, and Nullu across all models. This confirms its strength in reducing object-level hallucinations while maintaining descriptive relevance.

Results on POPE. We evaluate our method on the POPE benchmark (Li et al., 2023), which measures hallucination robustness under three query

Table 3: Evaluation Results on HallusionBench following the setting in (Guan et al., 2024).

Model	fACC	qACC	easyA	hardA	aACC
LLaVA-1.5-7B	17.9	8.1	36.0	36.7	41.5
Nullu	18.7	8.6	36.6	37.4	44.2
MPD	19.9	9.0	37.3	38.1	44.3
MiniGPT-4	10.1	8.7	31.8	27.6	35.7
Nullu	10.7	9.5	32.3	28.1	36.8
MPD	11.5	10.2	33.1	28.7	37.3
mPLUG-Owl2	19.9	13.8	44.8	39.1	47.3
Nullu	10.7	9.5	32.3	28.1	36.8
MPD	11.5	10.2	33.1	28.7	37.3

Table 4: Results on LLaVA-Bench following the setting in (Leng et al., 2024). Both metrics are on a scale of 10.

Model	Method	Accuracy	Detailedness
MiniGPT-4	Original	4.05	3.95
	MPD	**5.53**	**4.67**
mPLUG-Owl2	Original	5.76	4.22
	MPD	**6.13**	**4.62**
LLaVA-1.5-7B	Original	5.59	4.72
	MPD	**6.39**	**5.52**

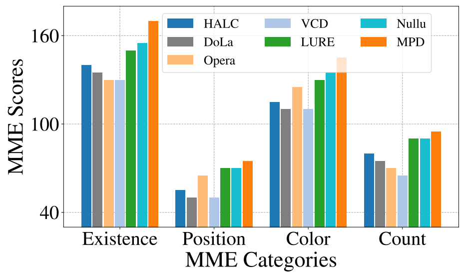

Figure 2: Comparisons of MME scores.

sampling strategies: *random*, *popular*, and *adversarial*. Table 2 reports results on LLaVA-1.5-7B, MiniGPT-4, and mPLUG-Owl2. We also report results on OPOPE (Li et al., 2023), AMBER (Wang et al., 2023), and MMHalBench (Sun et al., 2024). The results are provided in Appendix C.1.

- **Obs 3: MPD consistently achieves the highest F1 scores across all settings.** Under all sampling strategies, our method outperforms original models and the Nullu baseline. For example, in the *random* setting, we achieve 88.20 (LLaVA-1.5-7B), 85.76 (mPLUG-Owl2), and 71.81 (MiniGPT-4), with similar gains in *popular* and *adversarial* settings. These results confirm our approach's robustness and generalizability under varied hallucination conditions.
- **Obs 4: Complex visual scenes do not hinder the performance of MPD.** Even in adversarial settings with distractors, our method maintains high accuracy. On mPLUG-Owl2, we reach 78.72 F1 score, outperforming original (76.58) and Nullu (77.15), demonstrating effectiveness.

Results on Hallusionbench. HallusionBench dataset (Guan et al., 2024) evaluates fine-grained hallucinations beyond object existence. As shown in Table 3, MPDconsistently improves all metrics over the base models and Nullu across three LVLM backbones, including gains on challenging subsets such as *hardA* and *aACC*. These results indicate that our method generalizes beyond object-centric hallucinations. Further evidence on broader hallucination benchmarks is provided in Appendix C.1, as shown in Table 10.

4.3 Performance on Generative Capacity (RQ2)

To assess the general generative capacity of MPD beyond structured object grounding, we evaluate it on LLaVA-Bench and MME, which cover open-ended and structured settings (Leng et al., 2024). More results are provided in Appendix C.1.

Results on LLaVA-Bench. Table 4 reports GPT-4V-evaluated scores on *Accuracy* and *Detailedness*. MPD consistently improves both dimensions across all models. On LLaVA-1.5-7B, it raises accuracy from 5.59 to 6.39 and detailedness from 4.72 to 5.52, indicating enhanced factual alignment and richer responses without harming fluency.

- **Obs 5: Both grounding and response quality are improved by MPD.** The gains in both accuracy and detailedness suggest that MPD effectively reduces hallucinations while preserving the model's generative expressiveness.

Results on MME. Figure 2 summarizes results for four hallucination-sensitive subsets. MPD achieves the best performance on *Existence* and *Count*, measuring object hallucination, and remains competitive on *Position* and *Color*.

- **Obs 6: MPD strengthens robustness under structured hallucination evaluation.** The improvements on *Existence* and *Count* confirm that MPD enhances grounding accuracy in perception-oriented tasks, further supporting its generalization across evaluation paradigms.

Analysis of hidden representations. To assess the impact of editing on internal representations, we examine whether our method preserves the distributional structure of hidden features. We sample

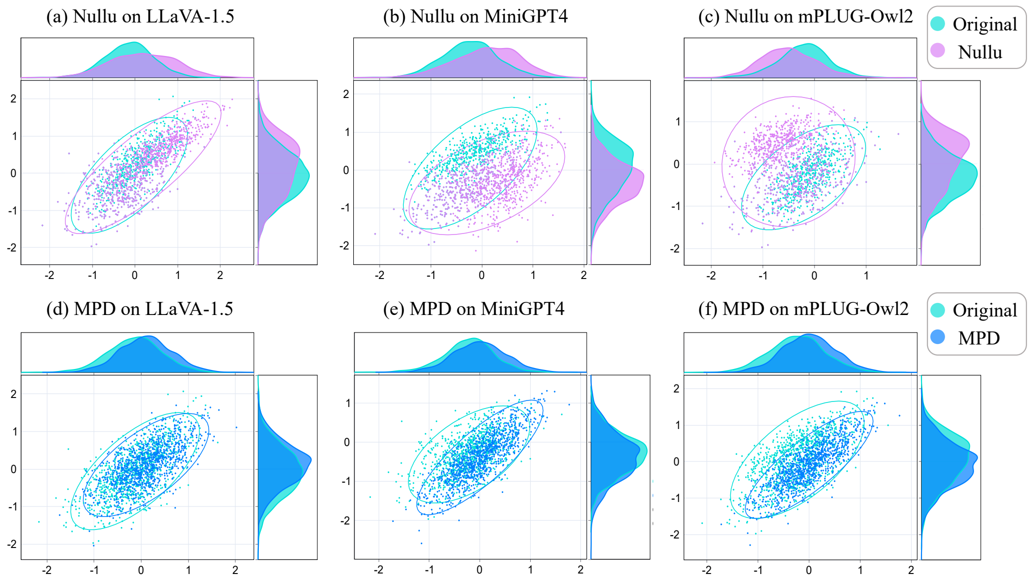

Figure 3: Distribution of hidden representations before and after editing across different LVLMs. Each subplot visualizes the principal components of token embeddings using PCA, comparing the original model and the edited version. The top and right marginal curves show the distributions along each principal axis. Compared to Nullu, our method (MPD) induces smaller representation shifts while achieving more effective hallucination suppression.

Table 5: Analysis of editing different layers.

$\{\ell\}$	CHAIR$_S$ ↓	CHAIR$_I$ ↓	BLEU↑
16-32	**12.80**	**4.20**	**15.31**
18-32	13.20	5.31	15.00
20-32	13.50	5.52	14.97
22-32	14.03	5.95	14.80
24-32	14.97	6.41	14.73
26-32	16.95	6.87	14.72
28-32	18.41	7.32	14.70
30-32	20.10	7.75	14.70

1,000 faithful prompts and extract their token-level representations from the original LVLMs. After applying edits via Nullu and MPD, we re-extract the features, project them into two dimensions using PCA, and visualize them along with their marginal distributions. As shown in Figure 3, MPD maintains close alignment with the original features, while Nullu introduces noticeable shifts.

4.4 Ablation Study (RQ3)

To analyze the contribution of each module in MPD, we perform ablation studies on editing layers and retained subspace dimensions using LLaVA-1.5-7B across CHAIR, POPE, and MME. More ablations are provided in Appendix C.2 and C.3.

Impact of editing layers. To assess the impact of editing layers, we perform an ablation study on LLaVA-1.5-7B using CHAIR$_S$, CHAIR$_I$, and BLEU metrics, as shown in Table 5.

- **Obs 7: Deeper edits yield hallucination suppression.** Editing from deeper layers (*e.g.*, 22–32 or 24–32) improves CHAIR scores, with the best result achieved at 16–32 (CHAIR$_S$ = 12.80, CHAIR$_I$ = 4.20). In contrast, narrow or shallow ranges (*e.g.*, 30–32 or 26–32) result in weaker performance and no BLEU gain, indicating insufficient removal of hallucinated components.

5 Conclusion

We present MPD, an optimization framework that mitigates hallucinations in LVLMs without degrading generative capacity. It disentangles hallucinatory components via orthogonal projection on contrastive representations and selectively updates parameters most correlated with these components using cosine similarity, thus avoiding large-scale perturbations. Extensive evaluations across multiple LVLMs and benchmarks show that MPD suppresses hallucinations while retaining 97.4% of generative capacity with no added inference cost.

Limitations

While MPD improves alignment on current benchmarks, it does not address hallucinations arising from broader data biases or prompt limitations. Moreover, automated benchmarks may not fully capture long-form coherence or stylistic diversity. Its robustness under highly ambiguous visual inputs remains to be evaluated.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (U24B20180, 62525211, 62576330) and the Natural Science Foundation of Anhui Province (2508085MF143).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2024a. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *NeurIPS*.

Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024b. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024c. Halc: Object hallucination reduction via adaptive focal-contrast decoding.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: towards general-purpose vision-language models with instruction tuning.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained knowledge editing for language models.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385. IEEE.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *AAAI*.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*.

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.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *CVPR*.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. 2025. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*.

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 *CVPR*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model.

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

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 *ECCV*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *NeurIPS*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *ACL (Findings)*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Rheeya Uppaal, Apratim Dey, Yiting He, Yiqiao Zhong, and Junjie Hu. 2024. Model editing as a robust and denoised variant of dpo: A case study on toxicity.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*.

Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, pages 13084–13094. IEEE.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *CVPR*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, pages 9556–9567. IEEE.

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

Ningyu Zhang, Bozhong Tian, Siyuan Cheng, Xiaozhuan Liang, Yi Hu, Kouying Xue, Yanjie Gou, Xi Chen, and Huajun Chen. 2024. Instructedit: Instruction-based knowledge editing for large language models. In *IJCAI*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*.

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

Xingyu Zhu, Beier Zhu, Junfeng Fang, Shuo Wang, Yin Zhang, Xiang Wang, and Xiangnan He. 2026b. Guardalign: Test-time safety alignment in multimodal large language models. *arXiv preprint arXiv:2602.24027*.

Xingyu Zhu, Beier Zhu, Shuo Wang, Junfeng Fang, Kesen Zhao, Hanwang Zhang, and Xiangnan He. 2026c. Principled steering via null-space projection for jailbreak defense in vision-language models. *arXiv preprint arXiv:2603.22094*.

Xingyu Zhu, Beier Zhu, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Enhancing clip robustness via cross-modality alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

A Proof of Proposition 1

We provide a detailed proof that the extracted hallucinatory component $\tilde{\mathbf{X}}_\ell = (\mathbf{I} - \mathbf{P}_\ell)\mathbf{X}_\ell^-$ is more accurate than the difference-based residual $\tilde{\mathbf{X}}_\ell^{\text{diff}} = \mathbf{X}_\ell^- - \mathbf{X}_\ell^+$ in estimating the orthogonal component of the hallucinated semantics $\mathbf{X}_\ell^{\text{hall},\perp}$, without assuming that the parallel component $\mathbf{X}_\ell^{\text{hall},\parallel} = \mathbf{P}_\ell \mathbf{X}_\ell^{\text{hall}}$ is small.

Given:

- $\mathbf{X}_\ell^+ \in \mathbb{R}^{N\times D}$: Faithful representations, spanning the faithful subspace.
- $\mathbf{X}_\ell^- = \mathbf{X}_\ell^{\text{real}} + \mathbf{X}_\ell^{\text{hall}} + \boldsymbol{\epsilon}^-$: Hallucinated representations, composed of real semantics, hallucinated components, and Gaussian noise $\boldsymbol{\epsilon}^- \sim \mathcal{N}(0, \sigma_-^2 \mathbf{I})$.
- $\mathbf{P}_\ell = \mathbf{U}_\ell \mathbf{U}_\ell^\top$: Orthogonal projection onto the subspace spanned by top-C singular vectors $\mathbf{U}_\ell$ from the SVD of $\mathbf{X}_\ell^+$.
- $\mathbf{X}_\ell^{\text{hall}} = \mathbf{X}_\ell^{\text{hall},\parallel} + \mathbf{X}_\ell^{\text{hall},\perp}$: Decomposition into in-subspace and orthogonal components w.r.t. $\mathbf{P}_\ell$.

The goal is to show:

$$\mathbb{E}\|\tilde{\mathbf{X}}_\ell - \mathbf{X}_\ell^{\text{hall},\perp}\|_F^2 \leq \mathbb{E}\|\tilde{\mathbf{X}}_\ell^{\text{diff}} - \mathbf{X}_\ell^{\text{hall},\perp}\|_F^2.$$

The projection-based residual is:

$$\tilde{\mathbf{X}}_\ell = (\mathbf{I}-\mathbf{P}_\ell)\mathbf{X}_\ell^- = (\mathbf{I}-\mathbf{P}_\ell)(\mathbf{X}_\ell^{\text{real}}+\mathbf{X}_\ell^{\text{hall}}+\boldsymbol{\epsilon}^-).$$

Decompose:

$$\tilde{\mathbf{X}}_\ell = (\mathbf{I}-\mathbf{P}_\ell)\mathbf{X}_\ell^{\text{real}} + (\mathbf{I}-\mathbf{P}_\ell)\mathbf{X}_\ell^{\text{hall}} + (\mathbf{I}-\mathbf{P}_\ell)\boldsymbol{\epsilon}^-.$$

Since $\mathbf{P}_\ell \mathbf{X}_\ell^{\text{real}} \approx \mathbf{X}_\ell^{\text{real}}$, we have $(\mathbf{I}-\mathbf{P}_\ell)\mathbf{X}_\ell^{\text{real}} \approx \mathbf{0}$. For the hallucinated semantics, $\mathbf{X}_\ell^{\text{hall}} = \mathbf{X}_\ell^{\text{hall},\parallel} + \mathbf{X}_\ell^{\text{hall},\perp}$, so:

$$(\mathbf{I}-\mathbf{P}_\ell)\mathbf{X}_\ell^{\text{hall}} = \mathbf{X}_\ell^{\text{hall},\perp},$$

since $(\mathbf{I}-\mathbf{P}_\ell)\mathbf{X}_\ell^{\text{hall},\parallel} = \mathbf{0}$. Thus:

$$\tilde{\mathbf{X}}_\ell \approx \mathbf{X}_\ell^{\text{hall},\perp} + (\mathbf{I}-\mathbf{P}_\ell)\boldsymbol{\epsilon}^-.$$

The error is:

$$\begin{aligned}\mathbb{E}\|\tilde{\mathbf{X}}_\ell - \mathbf{X}_\ell^{\text{hall},\perp}\|_F^2 &= \mathbb{E}\|(\mathbf{I}-\mathbf{P}_\ell)\boldsymbol{\epsilon}^-\|_F^2 \\ &= \sigma_-^2(D-C)N.\end{aligned}$$

since $\boldsymbol{\epsilon}^- \sim \mathcal{N}(0, \sigma_-^2\mathbf{I})$ and $\mathbf{I}-\mathbf{P}_\ell$ projects onto a $(D-C)$-dimensional subspace.

The difference-based residual is:

$$\begin{aligned}\tilde{\mathbf{X}}_\ell^{\text{diff}} &= \mathbf{X}_\ell^- - \mathbf{X}_\ell^+ \\ &= (\mathbf{X}_\ell^{\text{real}} + \mathbf{X}_\ell^{\text{hall}} + \boldsymbol{\epsilon}^-) - (\mathbf{X}_\ell^{\text{real}} + \boldsymbol{\epsilon}^+) \\ &= \mathbf{X}_\ell^{\text{hall}} + \boldsymbol{\epsilon}^- - \boldsymbol{\epsilon}^+.\end{aligned}$$

The error is:

$$\tilde{\mathbf{X}}_\ell^{\text{diff}} - \mathbf{X}_\ell^{\text{hall},\perp} = \mathbf{X}_\ell^{\text{hall},\parallel} + \boldsymbol{\epsilon}^- - \boldsymbol{\epsilon}^+.$$

The expected squared error is:

$$\mathbb{E}\|\tilde{\mathbf{X}}_\ell^{\text{diff}} - \mathbf{X}_\ell^{\text{hall},\perp}\|_F^2 = \|\mathbf{X}_\ell^{\text{hall},\parallel}\|_F^2 + \mathbb{E}\|\boldsymbol{\epsilon}^-\|_F^2 + \mathbb{E}\|\boldsymbol{\epsilon}^+\|_F^2,$$

since cross terms vanish under independence. Given $\mathbb{E}\|\boldsymbol{\epsilon}^-\|_F^2 = \sigma_-^2 DN$ and $\mathbb{E}\|\boldsymbol{\epsilon}^+\|_F^2 = \sigma_+^2 DN$:

$$\mathbb{E}\|\tilde{\mathbf{X}}_\ell^{\text{diff}} - \mathbf{X}_\ell^{\text{hall},\perp}\|_F^2 = \|\mathbf{X}_\ell^{\text{hall},\parallel}\|_F^2 + \sigma_-^2 DN + \sigma_+^2 DN.$$

Comparing the errors, the hallucinatory component has error $\sigma_-^2(D-C)N$, while the difference-based residual has error $\|\mathbf{X}_\ell^{\text{hall},\parallel}\|_F^2 + \sigma_-^2 DN + \sigma_+^2 DN$. Since $D - C < D$, we have $\sigma_-^2(D-C)N < \sigma_-^2 DN$. The additional terms $\|\mathbf{X}_\ell^{\text{hall},\parallel}\|_F^2 \geq 0$ and $\sigma_+^2 DN \geq 0$ ensure:

$$\sigma_-^2(D-C)N < \|\mathbf{X}_\ell^{\text{hall},\parallel}\|_F^2 + \sigma_-^2 DN + \sigma_+^2 DN.$$

The $\tilde{\mathbf{X}}_\ell$ isolates $\mathbf{X}_\ell^{\text{hall},\perp}$, corresponding to unique hallucinated concepts, and its error is unaffected by $\mathbf{X}_\ell^{\text{hall},\parallel}$, ensuring robustness.

B Implementation Details

We follow the experimental setup in previous work (Yang et al., 2025). During the decoding stage, beam search is employed with *num-beams* set to 3, indicating that three candidate sequences are retained at each step. For all models, the editing layers are selected within the range $\ell \in \text{range}(16, 32)$. The number of retained principal components is set to top-$C = 2500$, and the number of weights with the highest similarity scores is set to top-$K = 6000$. All experiments are conducted on a single A40 GPU (40G).

We construct faithful and hallucination-inducing contrastive pairs following the protocol in (Yang et al., 2025; Zhou et al., 2023). For each image and prompt, an auxiliary LLM is used to generate a faithful response grounded in the image and a hallucination-inducing response that introduces non-existent objects or attributes, and this construction is performed once offline before editing.

Table 6: The OPOPE evaluation results on MSCOCO dataset of LVLMs with different methods for mitigating OH. Higher accuracy, precision, and F1 score indicate better performance. Bold indicates the best result of all methods.

Method	MiniGPT-4			mPLUG-Owl2			LLaVA-1.5		
	Accuracy	Precision	F1 score	Accuracy	Precision	F1 score	Accuracy	Precision	F1 score
Greedy	$66.78_{\pm1.27}$	$90.43_{\pm25.1}$	$85.79_{\pm18.7}$	$69.77_{\pm1.18}$	$91.07_{\pm17.8}$	$87.45_{\pm13.9}$	$70.56_{\pm1.51}$	$91.08_{\pm20.6}$	$87.72_{\pm16.3}$
DoLA	$67.06_{\pm1.19}$	$90.84_{\pm23.1}$	$86.22_{\pm17.3}$	$70.17_{\pm1.69}$	$91.97_{\pm24.5}$	$88.30_{\pm19.2}$	$70.69_{\pm1.50}$	$90.87_{\pm19.8}$	$87.59_{\pm15.74}$
OPERA	$67.26_{\pm1.04}$	$90.76_{\pm20.0}$	$86.25_{\pm15.0}$	$69.26_{\pm0.45}$	$93.06_{\pm8.01}$	$88.83_{\pm6.14}$	$69.73_{\pm1.34}$	$91.10_{\pm19.4}$	$87.46_{\pm15.3}$
VCD	$65.78_{\pm0.96}$	$90.02_{\pm20.7}$	$85.00_{\pm15.1}$	$69.81_{\pm0.65}$	$92.70_{\pm11.0}$	$88.76_{\pm8.49}$	$70.67_{\pm1.22}$	$91.62_{\pm16.7}$	$88.19_{\pm13.3}$
LURE	$68.14_{\pm0.99}$	$90.95_{\pm17.34}$	$86.76_{\pm13.2}$	$69.24_{\pm1.60}$	$90.54_{\pm23.3}$	$86.85_{\pm18.2}$	$70.00_{\pm1.53}$	$90.89_{\pm21.9}$	$87.38_{\pm17.3}$
HALC	$66.76_{\pm0.68}$	$91.95_{\pm15.0}$	$86.92_{\pm11.1}$	$70.12_{\pm0.98}$	$91.94_{\pm15.1}$	$88.26_{\pm11.8}$	$70.59_{\pm0.82}$	$92.94_{\pm12.1}$	$89.22_{\pm9.55}$
Nullu	$68.81_{\pm0.59}$	$96.49_{\pm0.85}$	$91.21_{\pm0.65}$	$71.04_{\pm1.07}$	$96.30_{\pm0.39}$	$92.05_{\pm0.72}$	$72.52_{\pm0.14}$	$94.46_{\pm0.08}$	$92.79_{\pm0.14}$
MPD	$\mathbf{69.79}_{\pm0.60}$	$\mathbf{97.83}_{\pm0.16}$	$\mathbf{92.60}_{\pm0.08}$	$\mathbf{72.86}_{\pm1.00}$	$\mathbf{97.78}_{\pm0.03}$	$\mathbf{93.50}_{\pm0.02}$	$\mathbf{73.50}_{\pm0.45}$	$\mathbf{96.32}_{\pm0.05}$	$\mathbf{93.40}_{\pm0.07}$

Table 7: Comparison of hallucination mitigation methods in terms of object coverage (**Cover**) and hallucination rate (**HalRate**). Higher **Cover** and lower **HalRate** indicate better performance. Bold numbers denote the best results per model.

Method	MiniGPT-4		mPLUG-Owl2		LLaVA-1.5	
	Cover (↑)	HalRate (↓)	Cover (↑)	HalRate (↓)	Cover (↑)	HalRate (↓)
Greedy	65.2	69.2	53.5	41.8	50.5	26.4
VCD	52.1	46.1	63.2	70.9	50.9	25.3
HALC	62.6	65.9	52.7	39.2	49.9	25.9
Nullu	57.7	70.1	53.2	41.6	49.0	26.3
MPD	**63.2**	**45.7**	**53.4**	**38.5**	**51.2**	**25.1**

Table 8: Evaluation of different methods on hallucination severity and grounding quality. Higher **Score** and lower **HalRate** reflect stronger grounding. Bold values highlight the best performance for each model.

Method	MiniGPT-4		mPLUG-Owl2		LLaVA-1.5	
	Score (↑)	**HalRate** (↓)	**Score** (↑)	**HalRate** (↓)	**Score** (↑)	**HalRate** (↓)
Greedy	1.61	0.71	2.00	0.65	2.26	0.60
VCD	1.68	0.71	2.15	0.62	1.97	0.67
HALC	1.66	0.71	2.04	0.64	2.27	0.60
Nullu	1.32	0.77	2.19	0.61	1.61	0.71
MPD	**1.69**	**0.79**	**2.23**	**0.59**	**1.57**	**0.58**

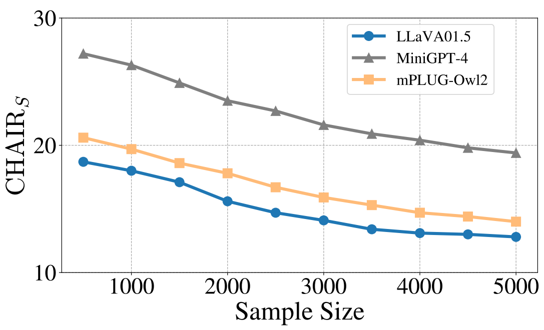

Figure 4: Comparisons of $CHAIR_S$ trends under varying the number of samples.

Table 9: Analysis of selected weights in model.

Values of K	POPE Acc.	$CHAIR_S$	$CHAIR_I$	MME
1×10^3	83.47	14.5	5.6	613.23
2×10^3	83.73	14.3	5.3	616.64
3×10^3	84.05	14.0	5.1	619.92
4×10^3	84.37	13.6	4.8	622.20
5×10^3	84.60	13.1	4.4	628.68
6×10^3	84.90	13.3	4.6	623.89
7×10^3	84.85	13.2	4.6	624.02
8×10^3	84.84	13.3	4.5	623.78

C Additional Results

C.1 Performance Evaluation

Results on OPOPE. To further validate our method, we evaluate on the OPOPE benchmark, which performs offline verification by checking whether generated captions mention presampled positive or negative objects. Compared to POPE (Li et al., 2023), OPOPE is more challenging as models must first describe the full image before grounding specific objects in text. Following standard protocol, we report Accuracy, Precision, and F_β (β=0.2) (Chen et al., 2024c) averaged across *random*, *popular*, and *adversarial* settings. Results are summarized in Table 6.

- **Obs: MPD consistently achieves the best or comparable performance across all models.** On LLaVA-1.5, MPD achieves the highest accuracy (73.50) and F1 score (93.40), outperforming Nullu by +0.98 and +0.61, respectively. On MiniGPT-4, we reach the best F1 score (92.60) and competitive accuracy (69.79). On mPLUG-Owl2, our method achieves the highest F1 score (93.50), surpassing Nullu and HALC. These results demonstrate that our method remains effective across models and metrics, improving object grounding accuracy while maintaining precision.

Results on AMBER. AMBER (Wang et al., 2023) is a generative evaluation benchmark that provides fine-grained object-level annotations across 1,000 images. It evaluates hallucination from multiple perspectives, including object coverage (Cover), hallucination rate (HalRate), and alignment with cognition. Table 7 reports object-level evaluation results on three LVLMs using AMBER metrics.

- **Obs: MPD achieves the best trade-off between coverage and hallucination suppression.** Compared with baselines, our method achieves the highest object coverage across all models (63.2 for MiniGPT-4, 53.4 for mPLUG-Owl2, and 51.2 for LLaVA-1.5), while also obtaining the lowest hallucination rates (45.7, 38.5, and 25.1, respectively). This suggests improved grounding performance without sacrificing content completeness.

Results on MMHalBench. MMHalBench (Sun et al., 2024) is a GPT-4-assisted question-answering benchmark that spans 12 object topics. Responses are rated by GPT-4 on a scale from 0 to 6, and those scoring below 3 are considered hallucinated. The metrics include average score and hallucination rate. Table 8 presents GPT-4 evaluation results following the MMHalBench protocol.

- **Obs: MPD improves response quality while reducing hallucinations.** Our method achieves the highest GPT-4 scores across all models (1.69 for MiniGPT-4, 2.23 for mPLUG-Owl2, and 1.57 for LLaVA-1.5), while maintaining the lowest hallucination rates (0.79, 0.59, and 0.58, respectively). These results indicate that our editing strategy produces more faithful and informative responses without degrading linguistic quality.

Table 10: Evaluation on non-object hallucination benchmarks.

Model	V* Bench			MMMU	MathVersion
	Attribute	Spatial	Overall		
LLaVA-1.5-7B	43.47	56.57	48.68	35.7	10.1
Nullu	45.08	57.26	49.15	35.9	10.3
MPD	47.73	58.91	51.03	36.5	11.0

Table 11: Effect of auxiliary LLM choice for contrastive pair construction.

Model	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	BLEU ↑
LLaVA-1.5-7B	20.40	7.08	15.72
MPD (GPT-3.5)	12.80	4.21	15.31
MPD (GPT-5.1)	12.60	4.21	15.02

Results on V* Bench, MMMU and MathVersion To examine generalization beyond object-centric hallucinations, we further evaluate MPD on V* Bench (Wu and Xie, 2024), which focuses on attribute and spatial consistency, as well as MMMU (Yue et al., 2024) and MathVision (Wang et al., 2024), which require multimodal reasoning over charts and abstract visual concepts. As shown in Table 10, MPDimproves the Attribute, Spatial, and Overall scores on V* Bench over both the base model and Nullu, indicating stronger fine-grained visual consistency. On MMMU and MathVision, although absolute performance remains modest due to task difficulty, MPD still yields consistent gains over the baselines, suggesting that our editing strategy does not harm (and can slightly benefit) broader multimodal reasoning ability.

Effect of Auxiliary LLM Choice. Our approach relies on an auxiliary LLM to construct faithful and hallucination contrastive pairs, motivating an examination of the LLM choice. Following prior representation intervention work, we use GPT-3.5 in our main experiments to ensure comparability. We then replace GPT-3.5 with GPT-5.1 and reconstruct all contrastive pairs under the same protocol. Table 11 shows that MPD yields consistent reductions in CHAIR_S and CHAIR_I and comparable BLEU scores across different auxiliary LLMs.

C.2 Impact of sample size.

We vary the sample size from 500 to 5000 and evaluate CHAIR$_S$ across three LVLMs. The results are shown in Figure 4.

- **Obs: More editing samples improve performance, while small-scale edits remain effective.** CHAIR$_S$ scores consistently decrease as the number of editing samples increases, indicating enhanced hallucination suppression. Notably, even with only 500 samples, our method yields clear gains over the original models, demonstrating strong data efficiency.

C.3 Effect of selected weights.

To understand how edited weight count affects hallucination suppression, we ablate top-K selection, where K is the number of weight rows aligned with hallucination semantics. Table 9 reports results on LLaVA-1.5-7B across POPE, CHAIR, and MME benchmarks.

- **Obs 10: Editing more weights improves performance up to a saturation point.** As K increases from 1×10^3 to 5×10^3, POPE accuracy steadily improves from 83.47 to 84.60, CHAIR$_S$/CHAIR$_I$ drop from 14.5/5.5 to 13.1/4.4, and MME score rises from 613.23 to 628.68. This consistently indicates that increasing the number of updated weights enhances hallucination suppression. However, further increasing K beyond 5×10^3 yields only marginal gains, suggesting that the generation of hallucinations are highly concentrated in a limited subset of parameters, and overly broad edits may lead to diminishing returns.

D Effiency Comparison.

Table 12: Efficiency and hallucination comparison across methods in terms of latency, GPU memory, and CHAIR$_S$. Experiments are evaluated on a single RTX A40 GPU.

Method	Avg. Latency ↓	GPU Memory ↓	CHAIR$_S$ ↓
Greedy	3.1 s	14758 MB	20.4
VCD	6.5 s	16538 MB	20.3
OPERA	23.5 s	23742 MB	17.5
HALC	20.1 s	22135 MB	16.9
Nullu	3.9 s	15183 MB	15.2
MPD	3.7 s	15019 MB	12.8

Table 12 presents a comparison of inference latency, peak GPU memory usage, and CHAIR$_S$ scores on the CHAIR benchmark using LLaVA-1.5 with a maximum token length of 128.

- **Obs: MPD achieves the best hallucination suppression with negligible runtime overhead.** While our method involves a lightweight parameter editing step prior to inference, it introduces only minimal latency increase compared to the greedy baseline (3.7s vs. 3.1s) and requires no additional modules during decoding. Moreover, it outperforms all baselines in CHAIR$_S$ (48.8), including optimization-heavy approaches such as HALC and OPERA, demonstrating an effective trade-off between performance and efficiency.

E Case Studies

Figure 5 shows case studies from LLaVA-Bench under identical prompts. The original model often hallucinates objects inconsistent with the image, such as *Mount Fuji*" for Diamond Head, or *fox*" and *musical instruments*" in a scene with no such elements. It also introduces non-existent food items like *broccoli*" and "*toast*". In contrast, our method generates responses that align with the visual content, avoiding co-occurrence-driven hallucinations.

Question: "What is the name of this famous sight in the photo?"

The image depicts Mount Fuji in Japan, known for its symmetrical shape and spiritual significance in Japanese culture. — Regular Response

The photo shows Diamond Head, a volcanic tuff cone on the Hawaiian island of O'ahu, near Honolulu. It is one of the most iconic landmarks in Hawaii. — Our Response

Question: "Describe this photo in detail."

The image shows passion fruits, one of which is sliced open to reveal yellow pulp and seeds. A spoon is placed nearby for serving. — Regular Response

The image shows several mangosteens on a wooden surface, with one cut open to reveal its white, segmented flesh. The fruits have dark purple rinds and green tops. — Our Response

Question: " What are the animals in the painting and what are they doing?"

The painting shows a bear, a fox, and a squirrel gathered around a picnic blanket, playing musical instruments in the forest. — Regular Response

The painting shows a bear, a cat, and a rabbit sitting around a tree stump, having tea and eating cookies together in a grassy field. — Our Response

Question: "What fruit is in the left part of the fridge?"

You can cook a vegetable stir-fry with broccoli and bell peppers, or make scrambled eggs with bacon and toast, using the items in the fridge. — Regular Response

You could prepare a snack plate with baby carrots and hummus, or a smoothie using milk, berries, and yogurt. — Our Response

Figure 5: Case studies on the LLaVA-Bench benchmark. We compare the responses generated by regular decoding and our method using LLaVA-1.5.