RFI: Rectified Flow Intervention for Mitigating Object Hallucination in Large Vision-Language Models

Junyu Cheng[1], Zhibiao Liang[2], Yidong Chen[1,3*], Shuangyin Li[2*]
[1]Department of Artificial Intelligence, School of Informatics, Xiamen University, China
[2]School of Computer Science, South China Normal University, China
[3]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China
junyucheng@stu.xmu.edu.cn, ydchen@xmu.edu.cn, shuangyinli@scnu.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in multimodal understanding and generation by integrating visual and textual data. However, these models frequently exhibit object hallucination problems: generating outputs that are inconsistent with the input image. Existing improved methods for mitigating hallucinations still suffer from two key limitations: dynamic approaches based on logits or attention mechanisms risk suppressing valuable linguistic priors, whereas static methods that employ fixed intervention vectors lack the flexibility to adapt to diverse images and questions. To address these issues, we propose **RFI** (Rectified Flow Intervention), a novel approach that harnesses the linear trajectory design of rectified flow for input-specific adaptation and employs gradient correction to ensure coherent generation, effectively combining the adaptability of dynamic methods with the stability of static ones. RFI dynamically predicts latent-space intervention vectors while requiring only a single forward pass in LVLMs per question, achieving computational efficiency (1.09x latency overhead for 100 new tokens). Extensive experiments show RFI significantly reduces hallucinations, achieving superior performance compared to existing advanced methods, highlighting its effectiveness as a lightweight plug-and-play method for reducing LVLM's hallucination in practical applications.

Code — https://github.com/purepasser-by/RFI.

Introduction

The rapid advancement of large vision-language models (LVLMs) (Liu et al. 2023; Zhu et al. 2023) has significantly enhanced multimodal understanding and generation capabilities. These models leverage cross-modal attention mechanisms and instruction-tuning techniques to generate contextually relevant responses.

Despite significant advancements in LVLMs, the persistent issue of hallucination (Zhu et al. 2024; Deng, Chen, and Hooi 2024; Liu et al. 2024c) continues to hinder their practical deployment. Hallucination manifests when LVLMs generate texts that misrepresent visual contents, such as mentioning non-existing objects, or fabricating object attributes.

*Corresponding author.

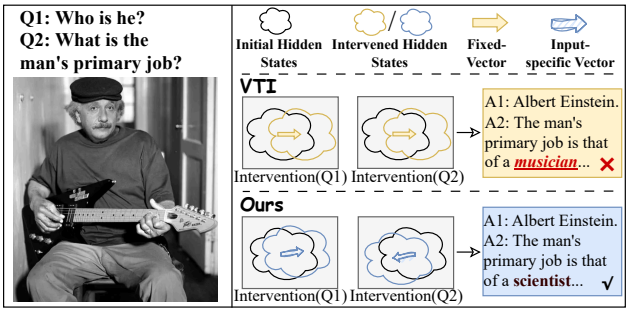

Figure 1: Comparison between VTI (Liu et al. 2024d) and RFI (Ours). VTI applies static intervention via fixed vectors, whereas RFI leverages rectified flow to predict input-specific intervention vectors. RFI dynamically adjusts hidden states while preserving essential language priors, effectively mitigating hallucinations in an context-sensitive manner.

Although linguistic priors inherited from LLMs can contribute to such errors, LVLM hallucinations can arise from driven by cross-modal misalignment, where visual inputs fail to adequately constrain generation. As shown in Figure 1, this is exemplified by an image of "Einstein playing guitar" leading the model to erroneously infer "Einstein was a musician," overgeneralizing from weak visual evidence under strong language priors.

Extensive research efforts have been devoted to hallucination mitigation, yielding a rich diversity of methodologies (Han et al. 2022; Wu et al. 2022; Chen et al. 2023; Yin et al. 2024; Sun et al. 2023; Liu et al. 2024a). We observe that prior approaches can be broadly categorized into two paradigms: dynamic and static interventions, each with distinct limitations in addressing the underlying causes of hallucination. **(1)** Dynamic approaches primarily address hallucination by mitigating linguistic prior dominance over visual signals, employing strategies like early-exit mechanisms (Zhou et al. 2024; Chuang et al. 2024) and visual contrastive decoding (Leng et al. 2023). However, these methods risk a fundamental trade-off: while reducing hallucination, they often yield overly conservative outputs (Huang et al. 2024) or fail to disentangle cross-modal misalignments (Favero et al. 2024), as visual and linguistic features remain entangled in high-dimensional space. Worsely, ex-

cessive suppression may introduce new hallucinations while impairing useful priors: a counterproductive "robbing Peter to pay Paul" effect (Chen et al. 2024). **(2)** Static intervention approaches, such as VTI (Liu et al. 2024d), ICT (Chen et al. 2024), and Nullu (Yang et al. 2025a), assume hallucination arises from systematic biases in joint vision-language feature space and precompute fixed correction vectors or weight edits to shift features away from hallucination-prone regions. While computationally efficient, these methods lack adaptability: they generalize poorly to novel scenarios (Gunjal, Yin, and Bas 2024), often under-correcting or over-penalizing semantic patterns. Critically, weight-editing may discard useful feature dimensions along with hallucinatory components (Yue, Zhang, and Jin 2024), and fixed vectors fail to generalize across diverse samples (Wang et al. 2025b). This limitation arises because static interventions inherently cannot address input-specific variations (as we analyze theoretically in Section Preliminary).

The shortcomings of existing approaches necessitate a solution that combines the precision of static methods with the adaptability of dynamic approaches. Rectified flow, which extends flow matching (Lipman et al. 2023) via building a linear trajectory between the source and the target, emerges as an ideal framework, addressing both requirements through its foundational properties. Where static methods falter by applying uniform interventions to diverse inputs, rectified flow's linear trajectory design (Liu, Gong, and Liu 2022) enables input-specific adaptation through input-conditioned path optimization. Conversely, while dynamic methods compromise output quality through aggressive feature suppression, rectified flow maintains generation coherence via gradient-aware flow matching (Lipman et al. 2023) that preserves linguistic integrity.

In light of these considerations, we propose a novel method named **RFI** (Rectified Flow Intervention), which innovatively incorporates the rectified flow model to dynamically generate intervention vectors tailored for each input pair of image and question. Specifically, RFI leverages paired hidden states from hallucinated and correct responses to train a rectified flow model, learning to transform the joint distribution of questions, images, and their erroneous latent representations toward fact-grounded answer spaces. During inference, this model dynamically predicts input-conditioned intervention vectors, which are then applied to the LVLM's hidden states to mitigate hallucinations and steer generation toward reliable outputs. As illustrated in Figure 1, while RFI generates input-specific intervention vectors to guide the model in latent space, VTI's static vectors remain susceptible to multimodal misalignment, leading to incorrect occupational inferences (e.g., concluding musician from playing guitar).

Our key contributions can be summarized as follows:

- We propose **RFI**, a novel method that leverages the rectified flow to dynamically predict intervention vectors for reducing hallucinations in LVLMs, capable of adapting to diverse input.
- We design a plug-and-play intervention framework that implements steering of latent space to dynamically correct hallucinations without fine-tuning the base LVLMs.
- Comprehensive experiments on LLaVA-v1.5 and Qwen-VL demonstrate that RFI significantly reduces hallucinations, outperforming existing intervention baselines.

Related Work

Large Vision-Language Models

The remarkable success of Large Language Models (LLMs) (Grattafiori, Dubey, and et al. 2024; Brown et al. 2020; Chiang et al. 2023; GLM et al. 2024; Bai et al. 2023a) has catalyzed significant advancements in Large Vision-Language Models (LVLMs). Early foundational works such as BLIP (Li et al. 2022, 2023a) and BERT-based VLMs (Devlin et al. 2019; Liu et al. 2019) demonstrated the feasibility of adapting LLMs for visual tasks through innovative architectural designs. Modern LVLMs typically employ a dual-encoder framework, combining visual encoders (e.g., CLIP (Radford et al. 2021)) with powerful LLMs through various fusion mechanisms, including linear projection layers and query-based transformers (Zhu et al. 2023). The introduction of visual instruction tuning (Liu et al. 2024b, 2023) has further enhanced these models' capabilities, enabling sophisticated image understanding and reasoning. This progress has yielded a new generation of high-performing LVLMs including LLaVA (Liu et al. 2024b), MiniGPT-4 (Zhu et al. 2023), mPLUG-Owl (Ye et al. 2024), and Qwen-VL (Bai et al. 2023b), which achieve state-of-the-art performance on multimodal benchmarks. However, these models inherit critical limitations from their LLM foundations, particularly the tendency to generate hallucinated content that misaligns with visual inputs (Jing and Du 2025; Bai et al. 2025; Wang et al. 2025c).

Hallucinations in LVLMs

Recent advancements in mitigating hallucinations within LVLMs can be broadly categorized into *dynamic* and *static* approaches.

Dynamic methods focus on real-time adjustments during inference, leveraging mechanisms such as attention steering, contrastive decoding, or iterative correction. For instance, (Wang et al. 2025d) introduces attention-steerable contrastive decoding to dynamically suppress hallucinatory outputs by contrasting candidate predictions. Similarly, (Li et al. 2025a) employs a training-free dynamic visual search to enhance fine-grained understanding without additional parameters. (Wang et al. 2025a) proposes dynamic correction decoding, where the model iteratively refines its predictions by revisiting visual inputs. (Park et al. 2025) mitigates perceptual hallucinations via selective and contrastive decoding, dynamically filtering implausible outputs. (Chen et al. 2024) performs cross-level trusted intervention during inference to correct object hallucinations. (Huang et al. 2024) penalizes over-trust in generated text through retrospection-allocation, dynamically adjusting confidence during response generation.

Static methods, in contrast, involve architectural modifications, latent space interventions, or training-based optimizations that operate independently of inference-time dy-

namics. (Yang et al. 2025a) projects hallucinatory outputs into a "HalluSpace" for offline mitigation, while (Liu et al. 2024d) employs latent space steering to align visual and textual representations. (Li et al. 2025b) reduces hallucinations by steering visual information at the token level. (Yang et al. 2025b) adopts modular attribution and intervention to decompose and rectify hallucinations. (Lyu et al. 2025) introduces hallucination-induced optimization during training. (Liu, Zheng, and Chen 2024) enhances visual reliance without fine-tuning by reweighting image-text alignment. (Wan et al. 2025) demonstrates that a single-layer intervention suffices to mitigate hallucinations, suggesting a lightweight static solution.

Preliminary

Limitations of Fixed Intervention

Fixed interventions inherently fail to address input-specific variations, which we now prove by demonstrating contradictory descent directions across samples.

Let $h_x \in \mathbb{R}^d$ be the hidden state produced by the model for an input x. Denote the deterministic downstream mapping by $g : \mathbb{R}^d \to \mathcal{Y}$. Hallucination is measured through a non–negative loss

$$\mathcal{L}(x, v) \;=\; \ell\big(g(h_x + v),\, y_x\big), \qquad v \in \mathbb{R}^d,$$

where y_x is the reference output that contains no hallucination. A *fixed intervention* adds the *same* vector v to all hidden states. We ask whether there exists a non–zero v such that

$$\mathcal{L}(x, v) \;\le\; \mathcal{L}(x, 0) \qquad \forall x. \tag{1}$$

First–order condition. A first–order Taylor expansion around $v = 0$ gives

$$\mathcal{L}(x, v) \;=\; \mathcal{L}(x, 0) + \underbrace{\big\langle \nabla_h \mathcal{L}(x, 0),\, v \big\rangle}_{=:\, \triangle(x,v)} + \mathcal{O}(\|v\|^2).$$

Write $g_x := \nabla_h \mathcal{L}(x, 0)$. Condition (1) requires

$$\langle g_x, v \rangle < 0 \qquad \forall x, \tag{2}$$

for sufficiently small $\|v\|$.

Proposition 1. *If there exist two samples x_1, x_2 such that $\langle g_{x_1},\, g_{x_2} \rangle < 0$, then no non–zero vector v satisfies* (2)*.*

Proof. Assume by contradiction that $v \neq 0$ obeys (2). Then $\langle g_{x_1}, v \rangle < 0$ and $\langle g_{x_2}, v \rangle < 0$, so their product is positive. However,

$$\begin{aligned} \langle g_{x_1}, v \rangle \, \langle g_{x_2}, v \rangle &= \|v\|^2 \big\langle g_{x_1}, \hat{v} \big\rangle \big\langle g_{x_2}, \hat{v} \big\rangle \\ &\le \tfrac{\|v\|^2}{2} \big(\|g_{x_1}\| \|g_{x_2}\| + \langle g_{x_1}, g_{x_2} \rangle \big) < 0, \end{aligned}$$

where $\hat{v} = v/\|v\|$ and the final inequality uses $\langle g_{x_1}, g_{x_2} \rangle < 0$. This contradicts the positivity of the product, proving the claim. □

Discussion. If the gradient covariance $\mathrm{Cov}[g_x]$ is full-rank (a typical scenario under a diverse data distribution), there must exist samples whose gradients are negatively correlated, activating Proposition 1. Consequently, a single fixed intervention can at most minimise *expected* hallucination loss, not guarantee per-sample reduction.

Rectified Flow

Rectified Flow (Liu 2022) is a flow matching model (Lipman et al. 2023) that constructs a linear trajectory between source and target distributions. Given a source distribution p_{source} and target distribution p_{target}, it learns a vector field that transports samples along straight paths with minimal curvature.

For paired samples $\mathbf{x} \sim p_{\text{source}}$ and $\mathbf{y} \sim p_{\text{target}}$, the linear interpolation path is defined as:

$$\mathbf{z}_t = t\mathbf{y} + (1 - t)\mathbf{x}, \quad t \in [0, 1], \tag{3}$$

where $\mathbf{z}_0 = \mathbf{x}$ and $\mathbf{z}_1 = \mathbf{y}$. The derivative of this path gives the ideal vector field direction:

$$\frac{\mathrm{d}\mathbf{z}_t}{\mathrm{d}t} = \mathbf{y} - \mathbf{x}, \tag{4}$$

The model trains a neural network $\mathbf{v}_\phi : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$ to approximate this ideal field by minimizing:

$$\min_\phi \int_0^1 \mathbb{E}_{\mathbf{x},\mathbf{y}} \left[\|(\mathbf{y} - \mathbf{x}) - \mathbf{v}_\phi(t, \mathbf{z}_t)\|_2^2 \right] \mathrm{d}t. \tag{5}$$

After training, new samples are generated by solving the ODE:

$$\mathrm{d}\mathbf{z}_t = \mathbf{v}_\phi(t, \mathbf{z}_t)\mathrm{d}t, \tag{6}$$

using numerical solvers (e.g., Euler (Euler 1845) or Runge-Kutta (Kutta 1901) methods).

Task Formulation

Given the textual input X, visual input V, a LVLM f_θ, and a rectified flow model $\mathcal{R}_\phi$, the concatenated input $q = concat(X, V)$.

After forwarding q, the hidden state of the last token at layer l (i.e., $\mathbf{h}_q^{(l)}$) is extracted. The hidden state is then passed to the rectified flow model, which outputs the intervention vector as:

$$\boldsymbol{\delta}^{(l)} = \mathcal{R}_\phi(\mathbf{h}_q^{(l)}). \tag{7}$$

The intervention vector $\boldsymbol{\delta}^{(l)}$ is purified via SVD through projection onto the top-k principal truth directions, yielding $\boldsymbol{\delta}_{\text{proj}}^{(l)}$. The purified vector is then injected with scaling factor α into the representation at decoder layer l:

$$\mathbf{h}_q^{(l)} \leftarrow \mathbf{h}_q^{(l)} + \alpha \boldsymbol{\delta}_{\text{proj}}^{(l)}. \tag{8}$$

Upon completion of autoregressive generation by the LLM, the original decoder layer representations are restored.

Methodology

Overview

As illustrated in Figure 2, our method comprises two key components: rectified flow training and intervention inference. (1) **Rectified Flow Training:** we first construct a dataset composed of positive-negative sample pairs to train the rectified flow model, enabling it to learn the trajectory from input queries to optimal intervention vectors. (2) **Intervention Inference:** given an input query, we initially forward it through the LVLM to extract its hidden states. The last token's hidden state is then fed into the pretrained rectified flow to sample the corresponding intervention vector. This vector is subsequently injected into the decoder's hidden layers to steer the generation process.

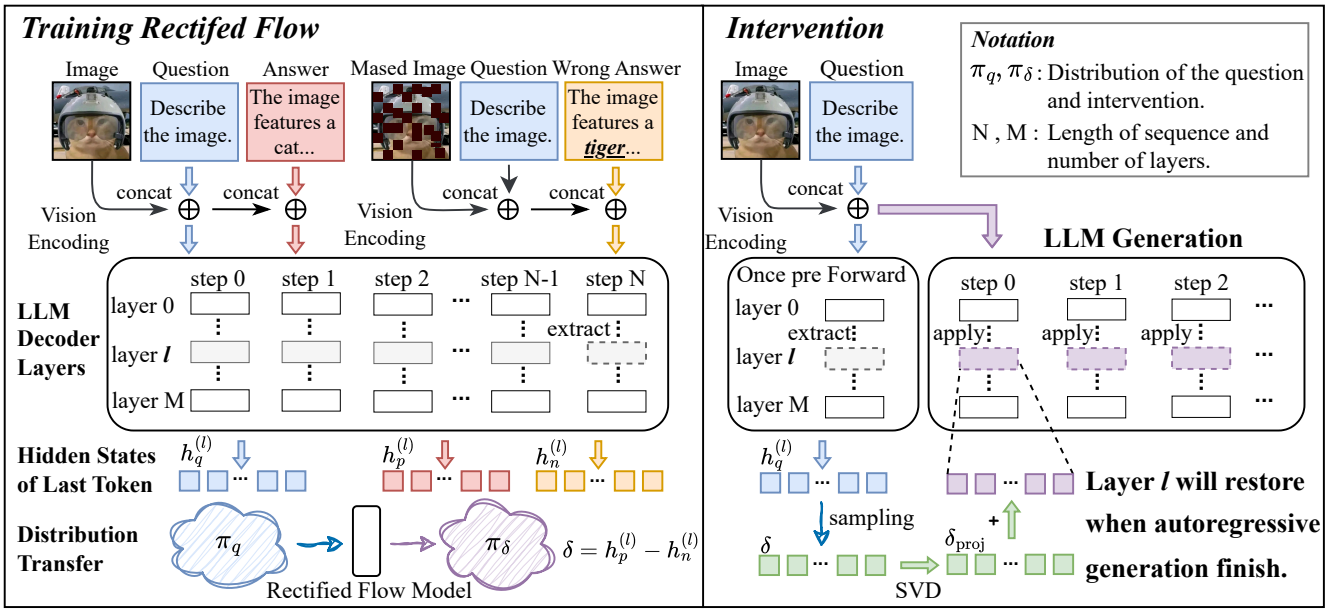

Figure 2: Overview of RFI. **(Left)** The training dataset for rectified flow is constructed using positive and negative sample pairs. Specifically, we concatenate (1) images with questions, (2) images with questions and correct answers, and (3) masked images with questions and hallucinated answers. These inputs are processed through LVLM to obtain the hidden states of the last token. From further computation, we obtain the source distribution q and target distribution δ for the rectified flow model. **(Right)** The LVLM's initial hidden states condition the trained rectified flow model to sample an intervention vector, which is then purified via singular value decomposition (SVD) and applied to modulate hidden states throughout generation.

Rectifed Flow Training

For a given visual input V, to generate the perturbed image $\tilde{V}$, we create m distinct randomly masked variants of V, each with a small black rectangular occlusion. The final $\tilde{V}$ is computed as the average of all masked versions.

For a given textual input X, concatenated input $q = concat(X, V)$ and $\tilde{q} = concat(X, \tilde{V})$, we consider a correct answer sample A_p and a hallucinated answer sample A_n. We then construct two sequences by concatenating the respective components, resulting in $S_p = concat(q, A_p)$ and $S_n = concat(\tilde{q}, A_n)$.

We first feed q into the LVLM and retrieve the hidden state of the last token at layer l, denoted as $\mathbf{h}_q^{(l)}$. Subsequently We feed S_p and S_n into the LVLM respectively and extract the hidden states at layer l corresponding to the last token. This process yields the hidden states $\mathbf{h}_p^{(l)}$ associated with the correct answer and $\mathbf{h}_n^{(l)}$ corresponding to the hallucinated answer. We then compute the intervention vector $\boldsymbol{\delta}^{(l)} = \mathbf{h}_p^{(l)} - \mathbf{h}_n^{(l)}$.

We have now obtained the key input-output variables $\mathbf{h}_q^{(l)}$ and $\boldsymbol{\delta}^{(l)}$ that are essential for training the rectified flow model.

Upon completing the construction of the dataset including collecting the concatenated input representation distribution π_q and the intervention vector distribution π_δ, we are set to commence the training process. The rectified flow training process for $\mathbf{h}_q^{(l)} \sim \pi_q$ and $\boldsymbol{\delta}^{(l)} \sim \pi_\delta$ follows:

$$\mathbf{z}_t = t\boldsymbol{\delta}^{(l)} + (1-t)\mathbf{h}_q^{(l)}, \tag{9}$$

which describes a linear interpolation between the query representation $\mathbf{h}_q^{(l)}$ and the intervention vector $\boldsymbol{\delta}^{(l)}$, where $t \in [0, 1]$ is a continuous parameter that governs the interpolation. For values of t in between, $\mathbf{z}_t$ traces a straight-line trajectory in the latent space, transitioning gradually from the source distribution π_q to the target distribution π_δ.

This deterministic path suggests a target direction for the velocity field $\mathbf{v}_{\boldsymbol{\phi}}(t, \mathbf{z}_t)$, which ideally should match the displacement $\boldsymbol{\delta}^{(l)} - \mathbf{h}_q^{(l)}$. To enforce this alignment, we minimize the expected squared error between the predicted velocity and the target displacement over all time steps and input pairs, resulting in the following loss functional:

$$\min_{\boldsymbol{\phi}} \int_0^1 \mathbb{E}_{\mathbf{h}_q^{(l)}, \boldsymbol{\delta}^{(l)} \sim \pi_q \otimes \pi_\delta} \left[\left\| \left(\boldsymbol{\delta}^{(l)} - \mathbf{h}_q^{(l)} \right) - \mathbf{v}_{\boldsymbol{\phi}}(t, \mathbf{z}_t) \right\|_2^2 \right] \mathrm{d}t. \tag{10}$$

Intervention Inference

After obtaining the trained rectified flow model, we can straightforwardly apply it to inference in LVLM. Given a test sample with textual input X, visual input V, and their concatenated input representation $q = concat(X, V)$, let f_θ denote the LVLM and $\mathcal{R}_\phi$ the rectified flow model. The rectified flow can be seamlessly integrated into the inference pipeline to refine the latent representations.

First, we feed the concatenated input q into the LVLM f_θ for a forward pass to obtain the hidden state of the last token at the l-th layer, denoted as $\mathbf{h}_q^{(l)}$. Subsequently, we perform flow sampling using the rectified flow model $\mathcal{R}_\phi$ to compute the initial intervention vector $\boldsymbol{\delta}^{(l)}$, formulated as $\boldsymbol{\delta}^{(l)} = \mathcal{R}_\phi(\mathbf{h}_q^{(l)})$.

However, the intervention vector $\boldsymbol{\delta}^{(l)}$ obtained from rectified flow sampling may contain noise and artifacts. To extract its most salient components, we employ Singular Value Decomposition (SVD) for denoising and projection. The core projection operation can be formalized as:

$$\boldsymbol{\delta}_{\text{proj}}^{(l)} = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^\top = \text{Proj}_k(\boldsymbol{\delta}^{(l)}), \tag{11}$$

where $\text{Proj}_k(\cdot)$ denotes the rank-k projection operator that preserves only the top-k singular components of the intervention vector. This projection effectively separates signal from noise by discarding smaller singular values that typically correspond to noisy or redundant dimensions in the latent space.

We subsequently define a wrapper module that encapsulates the hidden state processing pipeline, integrating the SVD. The core transformation is implemented as $\mathbf{h}_q^{(l)} \leftarrow \mathbf{h}_q^{(l)} + \alpha \boldsymbol{\delta}_{\text{proj}}^{(l)}$, where α serves as a scaling parameter that regulates the intervention intensity. This parametric formulation allows for balanced integration of the refined flow corrections while maintaining the stability of the original representations.

During the generation process, given a sequence length N and the current time step t (where $t \in \{1, \ldots, N\}$), let $\text{O}_t^{(l)}$ denote the output tuple of the replaced forward pass at the l-th layer. We formally define the modified output as follows:

$$\text{O}_t^{(l)} := (\mathbf{h}_t^{(l)} + \alpha \boldsymbol{\delta}_{\text{proj}}^{(l)}, \text{O}_{t,\setminus 0}^{(l)}), \tag{12}$$

where $\mathbf{h}_t^{(l)}$ represents the original hidden state at time step t. The notation $\text{O}_{t,\setminus 0}^{(l)}$ denotes all elements of the original output tuple O except the first (i.e., the hidden states).

Experiments

Experimental Settings

Datasets. **(1) POPE** (Li et al. 2023b) evaluates object hallucination in LVLMs using balanced Yes/No questions (50% existing/50% non-existing objects) from MSCOCO (Lin et al. 2015), A-OKVQA (Schwenk et al. 2022), and GQA (Hudson and Manning 2019). It employs standard VQA prompts (e.g., "*Is there a [object] in the image?*") to assess object recognition accuracy across 500 MSCOCO images. **(2) MME** (Fu et al. 2024) provides a comprehensive evaluation of LVLMs through 14 subtasks covering perceptual and cognitive abilities. It specifically measures hallucinations at object (existence, counting) and attribute (position, color) levels, using accuracy as the primary metric for standardized multimodal assessment.

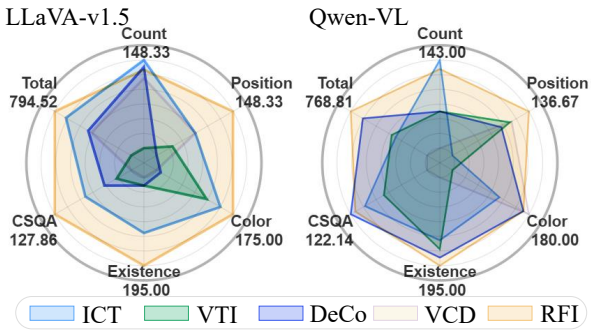

Figure 3: Comparison of RFI with baseline methods on the MME benchmark. The radar chart illustrates improvements across various evaluation categories, including existence, position, count, color, and commonsense QA (CSQA).

Baselines. We adopt the widely-implemented LLaVA-v1.5 (Liu et al. 2024b) and Qwen-VL (Bai et al. 2023b) architectures as our baseline LVLMs, both employing 7B-parameter language models. We compared four advanced baselines, including dynamic logit adjustment and fixed-vector intervention: **(1) VCD** (Leng et al. 2023), **(2) DeCo** (Wang et al. 2025a), **(3) VTI** (Liu et al. 2024d) and **(4) ICT** (Chen et al. 2024).

Implementation Details. We employed the dataset from (Liu et al. 2024d), which comprises 100 images from COCO (Lin et al. 2015) along with both correct and hallucinated captions. Then we performed random masking operations on these images to generate negative samples of images for our training corpus. For training the rectified flow model, we followed the methodology outlined in (Wang et al. 2025b; Lipman et al. 2023). Notably, the hidden states of different VLMs may affect the convergence behavior of the rectified flow model; accordingly, we train the rectified flow for 25 epochs on LLaVA-v1.5 and 40 epochs on Qwen-VL, respectively. All experiments were conducted on a single NVIDIA RTX A5000 GPU.

Main Results

Results on POPE. Table 1 presents the results of LLaVA-v1.5 and Qwen-VL on nine subsets of the POPE dataset. Our method yields two key conclusion: **(1)** Our approach achieves an average F1-score improvement of 7.59% (LLaVA-v1.5) and 6.90% (Qwen-VL), consistently outperforming all baselines. RFI demonstrates statistically significant improvements across all POPE subsets (random, popular, and adversarial), outperforming baseline methods in every category. This consistent performance gain confirms RFI's capability to generate effective and input-specific intervention vectors that adapt to diverse data distributions. **(2)** The rectified flow model attains robust generalization with merely 100 out-of-distribution sample pairs and seconds-level training time. This demonstrates its parameter-efficient nature while maintaining competitive intervention effectiveness.

Dataset	Setting	Category	Method	LLaVA-v1.5		Qwen-VL	
				Accuracy	F1 Score	Accuracy	F1 score
COCO	*Random*	–	Regular	83.29	81.33	84.37	82.67
		Dynamic	VCD	87.73	87.16	88.63	87.81
		Dynamic	DeCo	88.80	89.26	89.53	**89.31**
		Static	VTI	89.50	88.89	86.73	85.59
		Static	ICT	90.11	90.03	89.46	89.20
		–	RFI	**90.73** (↑ 7.44)	**90.40** (↑ 9.07)	**89.70** (↑ 5.33)	88.89 (↑ 6.22)
	Popular	–	Regular	81.88	80.06	84.13	82.06
		Dynamic	VCD	85.38	85.06	87.12	86.40
		Dynamic	DeCo	84.66	85.86	87.10	86.75
		Static	VTI	87.36	86.69	85.67	84.48
		Static	ICT	87.50	87.60	**88.16**	87.33
		–	RFI	**88.73** (↑ 6.85)	**88.21** (↑ 8.15)	87.93 (↑ 3.80)	**87.51** (↑ 5.45)
	Adversarial	–	Regular	78.96	77.57	82.26	80.37
		Dynamic	VCD	80.88	81.33	84.26	83.90
		Dynamic	DeCo	78.43	81.19	82.07	82.49
		Static	VTI	82.57	82.11	83.13	82.16
		Static	ICT	84.43	83.74	84.96	84.42
		–	RFI	**84.90** (↑ 5.94)	**84.68** (↑ 7.11)	**85.11** (↑ 2.85)	**84.49** (↑ 4.12)
AOKVQA	*Random*	–	Regular	83.45	82.56	86.67	85.59
		Dynamic	VCD	86.15	86.34	89.22	89.01
		Dynamic	DeCo	84.60	86.48	88.50	88.13
		Static	VTI	86.23	86.96	87.22	87.10
		Static	ICT	89.20	89.41	89.46	89.03
		–	RFI	**91.23** (↑ 7.78)	**91.15** (↑ 8.59)	**89.70** (↑ 3.03)	**89.30** (↑ 3.71)
	Popular	–	Regular	79.90	79.59	85.56	84.63
		Dynamic	VCD	81.85	82.82	87.85	87.81
		Dynamic	DeCo	80.40	81.34	88.23	87.89
		Static	VTI	80.66	82.61	87.94	87.56
		Static	ICT	85.73	85.34	88.13	87.83
		–	RFI	**87.57** (↑ 7.67)	**87.90** (↑ 8.31)	**90.23** (↑ 4.67)	**89.80** (↑ 5.17)
	Adversarial	–	Regular	74.04	75.15	79.57	79.50
		Dynamic	VCD	74.97	77.73	81.27	82.38
		Dynamic	DeCo	74.00	74.91	80.37	81.31
		Static	VTI	70.66	75.79	80.56	80.96
		Static	ICT	79.60	80.43	81.94	82.44
		–	RFI	**80.17** (↑ 6.13)	**82.00** (↑ 6.85)	**82.50** (↑ 2.93)	**83.09** (↑ 3.59)
GQA	*Random*	–	Regular	83.73	82.95	80.97	79.01
		Dynamic	VCD	86.65	86.99	85.59	85.33
		Dynamic	DeCo	84.80	86.67	79.03	78.81
		Static	VTI	87.10	87.94	85.89	85.17
		Static	ICT	89.60	89.44	86.38	86.96
		–	RFI	**89.73** (↑ 6.00)	**89.67** (↑ 6.72)	**90.23** (↑ 9.26)	**89.76** (↑ 10.75)
	Popular	–	Regular	78.17	78.37	75.99	74.84
		Dynamic	VCD	80.73	82.24	81.83	82.23
		Dynamic	DeCo	77.10	77.36	75.93	73.72
		Static	VTI	77.66	80.79	82.33	81.89
		Static	ICT	84.70	84.78	82.63	82.22
		–	RFI	**85.00** (↑ 6.83)	**85.30** (↑ 6.93)	**88.37** (↑ 12.38)	**88.04** (↑ 13.20)
	Adversarial	–	Regular	75.08	76.06	75.46	74.33
		Dynamic	VCD	76.09	78.78	80.01	80.75
		Dynamic	DeCo	73.13	74.47	75.13	74.68
		Static	VTI	72.33	77.24	80.46	80.09
		Static	ICT	81.50	82.27	80.83	80.60
		–	RFI	**81.90** (↑ 6.82)	**82.78** (↑ 6.72)	**83.97** (↑ 8.51)	**84.23** (↑ 9.90)

Table 1: Main results on POPE tasks. Baselines are categorized as *Static* or *Dynamic*. We evaluate accuracy and F1 Score across MSCOCO, A-OKVQA, and GQA datasets using LLaVA-v1.5 and Qwen-VL. Bold values indicate the best performance.

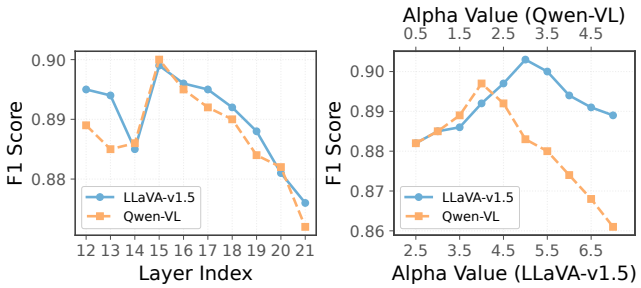

Figure 4: Ablation study of the hyper-parameters in RFI.

Results on MME. Figure 3 presents the results on the MME benchmark, where we following (Leng et al. 2023) evaluation step, focusing on the hallucination subset of MME. The experimental results demonstrate that RFI achieves consistent improvements across all evaluation metrics, outperforming the baseline method ICT by 27.91 points on LLaVA-v1.5 and 47.48 points on Qwen-VL. RFI exhibits substantial performance gains on non-hallucination metrics for both model architectures, indicating its dual capability in effectively mitigating visual hallucinations while simultaneously enhancing general multimodal reasoning performance. Our method shows slightly lower performance than ICT in counting tasks, likely due to the latter's finer-grained object-level method.

Ablations and Analysis

Effects of Layer l and Intensity α. RFI contains two key parameters: the target layer index l requiring intervention and the intervention intensity α. We evaluate these hyperparameters with LLaVA-v1.5 and Qwen-VL on the random subset of COCO from the POPE benchmark. To investigate the impact of these parameters, we fixed one of the parameters and calculate the F1 score. Figure 4 (left) demonstrates that both models achieve peak performance at approximately layer 15, with performance degradation observed when deviating from this optimal depth. This phenomenon suggests layer 15 serves as a critical fusion point for cross-modal integration, where earlier layers may lack sufficient feature abstraction while deeper layers could suffer from over-fusion or information distortion. Figure 4 (right) illustrates the performance across α values. Empirical findings show that the suppression of hallucinations is maximized when α values approach 5.0 in LLaVA-v1.5 and 2.0 in Qwen-VL, suggesting greater parameter sensitivity in the latter model.

Inference Latency. Table 2 compares the inference latency of LLaVA-v1.5 before and after applying our RFI method across varying new token lengths, where each configuration was evaluated through 20 independent trials. Remarkably, at 100 new tokens, RFI introduces only a 1.09x overhead, demonstrating the exceptional efficiency of our approach. This lightweight overhead is achieved through just one additional forward pass for hidden state extraction and a single rectified flow sampling operation.

New Tokens	LLaVA-v1.5	+ RFI	Ratio
5	0.28 ± 0.005	0.53 ± 0.005	1.89x
25	0.81 ± 0.007	1.07 ± 0.007	1.32x
50	1.47 ± 0.006	1.73 ± 0.007	1.17x
75	2.14 ± 0.014	2.40 ± 0.031	1.12x
100	2.79 ± 0.015	3.05 ± 0.020	1.09x

Table 2: Comparison of the latency (s) of LLaVA-v1.5 before and after applying RFI in generating tokens of varying lengths on an NVIDIA A5000 GPU. "±" represents the standard deviation.

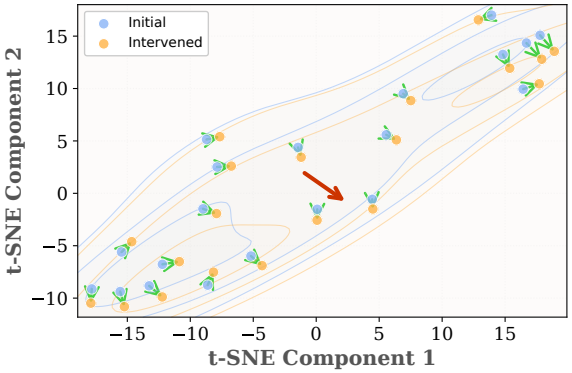

Figure 5: Visualization of initial and intervened hidden states in LLaVA-v1.5 at layer 15 using t-SNE.

Analysis of Input-Specific Intervention. As shown in Figure 5, the visualization of 20 randomly sampled COCO instances demonstrates that the input-specific intervention vectors induce systematic and directional shifts in the hidden states of LLaVA-v1.5 at the 15-th layer, as evidenced by the coherent displacement patterns between the initial (blue) and intervened (orange) points and the mean intervention direction (red) in the t-SNE space. The locally concentrated arrows in high-density regions suggest targeted modifications to task-relevant features, while globally preserved dispersion indicates maintained representational diversity without collapse. While the mean intervention (red) demonstrates the dominant transformation pattern, individual cases exhibit significant directional variation. This contrasts with static intervention methods, as RFI dynamically adapts to input content, achieving fine-grained control. Crucially, the coexistence of aligned and opposed interventions implies our method captures nuanced error-specific corrections beyond fixed intervention.

Conclusion

We propose RFI, a novel method that leverages rectified flow to dynamically predict latent-space intervention vectors. Requiring only a single forward pass in LVLMs per question, RFI achieves computational efficiency. Extensive experiments demonstrate that RFI significantly outperforms existing baselines in hallucination reduction.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grants 62476232, 62566060) and the University-Industry Cooperation Program of Fujian Province (Grant 2023H6001).

References

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; and et al., K. D. 2023a. Qwen Technical Report. arXiv:2309.16609.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023b. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2025. Hallucination of Multimodal Large Language Models: A Survey. arXiv:2404.18930.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, J.; Zhang, T.; Huang, S.; Niu, Y.; Zhang, L.; Wen, L.; and Hu, X. 2024. ICT: Image-Object Cross-Level Trusted Intervention for Mitigating Object Hallucination in Large Vision-Language Models. arXiv:2411.15268.

Chen, Z.; Zhu, Y.; Zhan, Y.; Li, Z.; Zhao, C.; Wang, J.; and Tang, M. 2023. Mitigating Hallucination in Visual Language Models with Visual Supervision. arXiv:2311.16479.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; and He, P. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. arXiv:2309.03883.

Deng, A.; Chen, Z.; and Hooi, B. 2024. Seeing is Believing: Mitigating Hallucination in Large Vision-Language Models via CLIP-Guided Decoding. arXiv:2402.15300.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Euler, L. 1845. *Institutionum calculi integralis*. Institutionum calculi integralis. Impensis Academiae Imperialis Scientiarum.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multi-Modal Hallucination Control by Visual Information Grounding. arXiv:2403.14003.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

GLM, T.; :; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; and et al., D. Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Grattafiori, A.; Dubey, A.; and et al., A. J. 2024. The Llama 3 Herd of Models. arXiv:2407.12345.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and Preventing Hallucinations in Large Vision Language Models. arXiv:2308.06394.

Han, Y.; Nie, L.; Yin, J.; Wu, J.; and Yan, Y. 2022. Visual Perturbation-aware Collaborative Learning for Overcoming the Language Prior Problem. arXiv:2207.11850.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. arXiv:2311.17911.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. arXiv:1902.09506.

Jing, L.; and Du, X. 2025. FGAIF: Aligning Large Vision-Language Models with Fine-grained AI Feedback. arXiv:2404.05046.

Kutta, W. 1901. *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen, Inaugural-Dissertation ... von Wilhelm Kutta ...* Druck von B.G. Teubner.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2023. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. arXiv:2311.16922.

Li, G.; Xu, J.; Zhao, Y.; and Peng, Y. 2025a. DyFo: A Training-Free Dynamic Focus Visual Search for Enhancing LMMs in Fine-Grained Visual Understanding. arXiv:2504.14920.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating Object Hallucination in Large Vision-Language Models. arXiv:2305.10355.

Li, Z.; Shi, H.; Gao, Y.; Liu, D.; Wang, Z.; Chen, Y.; Liu, T.; Zhao, L.; Wang, H.; and Metaxas, D. N. 2025b. The Hidden Life of Tokens: Reducing Hallucination of Large Vision-Language Models via Visual Information Steering. arXiv:2502.03628.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. arXiv:2210.02747.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2024a. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. arXiv:2306.14565.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024b. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024c. A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253.

Liu, Q. 2022. Rectified Flow: A Marginal Preserving Approach to Optimal Transport. arXiv:2209.14577.

Liu, S.; Ye, H.; Xing, L.; and Zou, J. 2024d. Reducing Hallucinations in Vision-Language Models via Latent Space Steering. arXiv:2410.15778.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. arXiv:2407.21771.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. arXiv:2209.03003.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Lyu, X.; Chen, B.; Gao, L.; Song, J.; and Shen, H. T. 2025. Alleviating Hallucinations in Large Vision-Language Models through Hallucination-Induced Optimization. arXiv:2405.15356.

Park, W.; Kim, W.; Kim, J.; and Do, J. 2025. SECOND: Mitigating Perceptual Hallucination in Vision-Language Models via Selective and Contrastive Decoding. arXiv:2506.08391.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. arXiv:2206.01718.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; Keutzer, K.; and Darrell, T. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. arXiv:2309.14525.

Wan, Z.; Zhang, C.; Yong, S.; Ma, M. Q.; Stepputtis, S.; Morency, L.-P.; Ramanan, D.; Sycara, K.; and Xie, Y. 2025. ONLY: One-Layer Intervention Sufficiently Mitigates Hallucinations in Large Vision-Language Models. arXiv:2507.00898.

Wang, C.; Chen, X.; Zhang, N.; Tian, B.; Xu, H.; Deng, S.; and Chen, H. 2025a. MLLM can see? Dynamic Correction Decoding for Hallucination Mitigation. arXiv:2410.11779.

Wang, H.; Cao, B.; Cao, Y.; and Chen, J. 2025b. TruthFlow: Truthful LLM Generation via Representation Flow Correction. arXiv:2502.04556.

Wang, H.; Yue, Y.; Lu, R.; Shi, J.; Zhao, A.; Wang, S.; Song, S.; and Huang, G. 2025c. Model Surgery: Modulating LLM's Behavior Via Simple Parameter Editing. arXiv:2407.08770.

Wang, Y.; Bi, J.; Ma, Y.; and Pirk, S. 2025d. ASCD: Attention-Steerable Contrastive Decoding for Reducing Hallucination in MLLM. arXiv:2506.14766.

Wu, Y.; Zhao, Y.; Zhao, S.; Zhang, Y.; Yuan, X.; Zhao, G.; and Jiang, N. 2022. Overcoming Language Priors in Visual Question Answering via Distinguishing Superficially Similar Instances. arXiv:2209.08529.

Yang, L.; Zheng, Z.; Chen, B.; Zhao, Z.; Lin, C.; and Shen, C. 2025a. Nullu: Mitigating Object Hallucinations in Large Vision-Language Models via HalluSpace Projection. arXiv:2412.13817.

Yang, T.; Li, Z.; Cao, J.; and Xu, C. 2025b. Mitigating Hallucination in Large Vision-Language Models via Modular Attribution and Intervention. In *The Thirteenth International Conference on Learning Representations*.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. arXiv:2408.04840.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2024. Woodpecker: hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12).

Yue, Z.; Zhang, L.; and Jin, Q. 2024. Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective. arXiv:2402.14545.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. arXiv:2310.00754.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.

Zhu, L.; Ji, D.; Chen, T.; Xu, P.; Ye, J.; and Liu, J. 2024. IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding. arXiv:2402.18476.