DARE to Mitigate Hallucination: Dual-path Auto-Regressive-aware Editing

Jae-Ho Lee, Jeong-Eun Lee, and Gyeong-Moon Park†

Korea University, Seoul, Republic of Korea
{jaeho-lee, esilver, gm-park}@korea.ac.kr

Abstract. Large vision-language models (LVLMs) have recently achieved remarkable progress across multimodal tasks, yet object hallucination remains a persistent challenge where models generate descriptions inconsistent with the visual input. Recent work mitigates hallucinations through training-free representation editing, typically by constructing hallucination-related directions from teacher-forcing (TF) contrasts between hallucinated and truthful responses. However, LVLMs operate through autoregressive (AR) decoding during generation, raising the question of whether TF-based analysis fully reflects the generation dynamics that lead to hallucinated outputs. In this paper, we analyze the relationship between TF-based editing and AR generation behavior and find that TF-based editing alone may be insufficient to capture both decoding dynamics and multimodal interactions associated with hallucinations. To address this limitation, we propose **DARE (Dual-path Auto-Regressive-aware Editing)**, a hybrid hallucination editing framework that integrates two complementary contrast pathways: textual contrasts and image contrasts, together with autoregressive-aware representation signals. Specifically, DARE constructs hallucination editing directions from (1) TF-based textual contrasts, (2) AR-aware representation transitions during decoding, and (3) controlled visual differences between paired images. Extensive experiments on multiple LVLM hallucination benchmarks demonstrate that DARE consistently reduces object hallucinations while preserving multimodal perception capability and inference efficiency. Our implementation code is available at https://github.com/KU-VGI/DARE.

Keywords: Large Vision-Language Models · Hallucination Mitigation · Representation Editing

1 Introduction

Large Vision-Language Models (LVLMs) [35,57,64] have recently achieved remarkable progress across a wide range of multimodal tasks, including visual question answering, image captioning, and multimodal dialogue. Despite these advances, **Object Hallucination (OH)** [22,24,31,33,34,45] remains a persistent

† Corresponding Author.

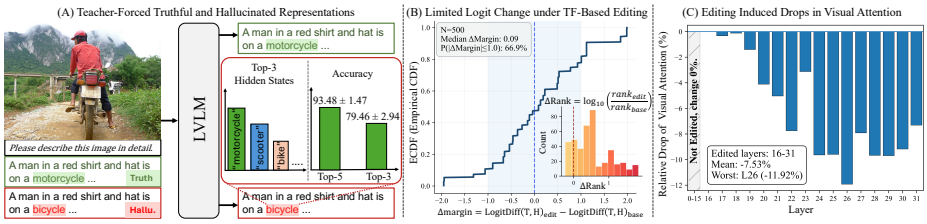

Fig. 1: Limitations of teacher-forcing (TF)-based hallucination editing in LVLMs. (A) Under TF evaluation, truthful tokens frequently remain among the top-ranked candidates even when evaluated on hallucinated captions under TF. (B) TF-based editing modifies token distributions but results in limited changes in the logit margin between truthful and hallucinated tokens, indicating that generation decisions often remain unchanged. The inset histogram further shows substantial shifts in token rankings. (C) Editing directions derived from textual contrasts are accompanied by reduced attention to visual tokens across layers. These observations highlight limitations of TF-based hallucination editing and motivate approaches that explicitly consider generation dynamics and multimodal representations.

challenge, where models generate textual content that is inconsistent with objects present in the visual input. Such hallucinations reduce the reliability of LVLMs and limit their deployment in real-world applications [10,11,19,41]. Consequently, a growing body of research [15,20,21,28,53,55,59] has focused on understanding the causes of hallucinations and developing techniques to mitigate them.

Among recent approaches, representation editing [23,55,56] has emerged as a practical strategy for hallucination mitigation. Instead of retraining [18,61,62], these methods perform *training-free weight editing* by identifying directions associated with hallucinated responses. In particular, several approaches [14,55] construct a hallucination-related subspace (hereafter HalluSpace) by contrasting truthful and hallucinated responses under teacher-forcing (TF). The model is then edited through null-space projection that removes hallucination-associated components from the representation space. This approach enables hallucination mitigation without additional training while preserving inference efficiency.

To better understand how TF-based editing operates, we first analyze token distributions under TF evaluation. Surprisingly, we observe that truthful tokens often remain highly ranked under TF evaluation, even when conditioning on hallucinated captions. As illustrated in Fig. 1(A), the ground-truth token frequently appears within the top-ranked candidates of hallucinated representations. In particular, we measure how often the truthful token appears among the top-k candidates of the hidden representation, and find that it remains within the top-5 candidates with high probability. This observation indicates that the model often retains a strong internal preference toward the correct token even in hallucination cases. Therefore, hallucinated generations may arise not from the absence of correct knowledge in the representation, but from how relative token scores are compared and selected during auto-regressive generation.

Motivated by this observation, we investigate whether TF-based editing actually changes the token selection during auto-regressive generation. If TF-based editing effectively reduces hallucinations, it should shift the model's preference from the hallucinated token to the truthful token at decoding time. To test this, we compute the logit margin between the truthful token and the hallucinated token before and after editing on MSCOCO [32]. Fig. 1(B) shows the empirical cumulative distribution function (ECDF) of the margin difference between the edited and base models for hallucinated samples generated under auto-regressive decoding. We find that most margin differences are close to zero, meaning that TF-based editing only slightly changes the relative scores of the two tokens. Since autoregressive decoding selects the token with the highest logit, the generated output changes only if the truthful token overtakes the hallucinated token in rank. However, when the margin remains nearly unchanged, the hallucinated token often continues to have the higher score and is therefore still selected. This analysis suggests that although TF-based editing can modify hidden representations and slightly adjust token distributions, it often fails to change the final decoding decision. As a result, hallucinations may persist even when the truthful token is ranked highly in the model's internal representation.

Beyond generation dynamics, we examine whether hallucination directions derived from *text-only* contrasts can inadvertently affect multimodal interactions in LVLMs. Most prior methods [14, 23, 55, 56] construct editing directions by contrasting hallucinated and truthful textual responses, implicitly treating hallucination as primarily a linguistic difference. However, Fig. 1 (A) shows that even for hallucinated evaluation, the model often retains strong evidence for the truthful token under teacher-forcing, suggesting that image-conditioned information remains present in the internal representation. Since such representations are already multimodal and entangled, text-only editing directions may correlate with components that also support visual-token interactions. Consistently, Fig. 1 (C) shows that applying text-only editing is accompanied by reduced attention to visual tokens across layers, indicating a potential collateral effect on visual pathways when editing is defined solely from textual contrasts.

Based on these observations, we propose a hybrid editing framework coined **DARE (Dual-path Auto-Regressive-aware Editing)** designed to address both limitations. DARE is built on two key principles: *dual-path contrasts* and *autoregressive-aware editing*. First, DARE constructs hallucination editing directions through **two complementary contrast pathways**. The first pathway exploits **textual contrasts** between hallucinated and truthful responses under TF evaluation, capturing linguistic representation differences associated with hallucination. The second pathway utilizes **image contrasts** (via **Visual-Difference Editing**), where controlled visual differences between paired images induce representation changes grounded in visual evidence. DARE integrates textual and AR-aware directions into an **Augmented HalluSpace** and complements it with **Visual-Difference Editing** based on image contrasts. Empirically, DARE consistently reduces hallucination rates across multiple LVLM benchmarks while maintaining generation quality. Further analyses show improved

stability of generation decisions and more consistent multimodal behaviors during autoregressive generation. Our contributions are summarized as follows:

- We provide a systematic analysis of TF-based hallucination editing methods and show that (i) truthful tokens often remain highly ranked under TF evaluation even when hallucinations occur, (ii) despite reshaping token distributions, TF-based editing frequently yields only limited changes in token-level decoding decisions under AR generation, and (iii) text-only editing directions can be accompanied by reduced attention to visual tokens across layers.
- Based on these analyses, we propose **DARE (Dual-path Auto-Regressive-aware Editing)**, a training-free weight editing framework that integrates **dual-path contrasts**, **text contrasts** and **image contrasts** and further incorporates **AR-aware signals** derived from AR decoding trajectories to better target hallucination-inducing behaviors.
- We conducted extensive experiments on multiple LVLM hallucination benchmarks, demonstrating that DARE consistently reduces object hallucinations while preserving multimodal perception capability and inference efficiency.

2 Related Work

Large Vision-Language Models. Large Vision-Language Models (LVLMs) have rapidly evolved following the success of Large Language Models (LLMs). This progress, in turn, has led to significant improvements in multimodal tasks such as image captioning [1,52,60] and visual question answering [2,50]. Specifically, to further enhance multimodal understanding, LLaVA [36] and LLaVA-1.5 [35] integrate pretrained vision encoders with language decoders and employ instruction tuning [42] to improve visual reasoning and comprehension. MiniGPT-4 [64] utilizes a Q-former [29] to efficiently fuse visual and textual features while reducing redundant visual tokens. mPLUG-Owl2 [57] adopts a parameter-efficient adaptation strategy that trains image-conditioned visual prompts instead of performing full model fine-tuning. Beyond these representative architectures, more advanced line of LVLM models [4,7,9,25,29] have emerged.

Mitigating Object Hallucination in LVLMs. Object hallucination, which refers to the inconsistency between the objects in visual content and the generated text description [46], has been a persistent challenge in LVLMs. Approaches for mitigating hallucinations can generally be categorized according to the stage at which they are applied. The first line of research targets the training stage, where models are improved with newly optimized training objectives [13,40,47] or through the use of task-specifically constructed datasets [39,43] for OH mitigation. Although these approaches have demonstrated promising results, they require extensive computational resources and time. Naturally, another line of research has emerged, that focuses on inference-stage techniques including advanced decoding–based strategies [8,12,20,21,28,30,53] as well as attention intervention methods [38,59]. In addition, some recent works have attempted to mitigate hallucinations by editing model weights or steering the latent space. While latent

steering methods [26,27,37,49] adjust latent features by adding precomputed stable directions at inference time, weight-editing approaches [14,55] project model weights onto the null space of hallucination-related components to explicitly attenuate them. Our work addresses hallucination by refining the weights of LVLMs, but in a way that more effectively preserves multimodal dependencies.

Teacher-forcing and Auto-regressive Generation Gap. The discrepancy between teacher-forcing (TF) and auto-regressive (AR) generation has long been studied in language sequence modeling as the exposure bias problem [3,6,44,48,54]. During TF, models are conditioned on ground-truth tokens, whereas during AR they rely on their own previously generated outputs. This mismatch can lead to cascading errors and unstable generation dynamics. While exposure bias has been extensively studied in language sequence modeling, its implications for hallucinated generation in LVLMs remain unclear. In particular, several hallucination editing methods derive editing directions by analyzing hidden representations under TF settings. However, hallucinated outputs arise during AR generation, where token selection depends on relative score comparisons among candidate tokens. This discrepancy suggests that editing directions derived from teacher-forced representations may not fully reflect the generation dynamics that influence token selection during generation. Our work revisits this gap and analyzes its implications for hallucination mitigation in LVLMs.

3 Method

This section introduces **DARE (Dual-path Auto-Regressive-aware Editing)**, a hallucination mitigation method for LVLMs, as illustrated in Fig. 2. DARE constructs hallucination editing directions through two complementary contrast pathways together with autoregressive-aware signals. Specifically, DARE leverages (1) *textual contrasts* derived from hallucinated and truthful responses under teacher-forcing, and (2) *image contrasts* obtained from controlled visual differences. In addition, DARE incorporates *Auto-Regressive-aware* signals that explicitly capture representation transitions occurring during decoding. Together, these components enable DARE to identify hallucination-related representation directions while preserving visually grounded multimodal interactions.

We first review TF-based hallucination editing methods that construct hallucination subspaces from textual representation contrasts. We then introduce the *AR-aware HalluSpace*, which derives hallucination directions from autoregressive generation trajectories. Finally, we present the full DARE framework that integrates textual contrasts, autoregressive dynamics, and visual contrasts.

3.1 Preliminary: TF-based HalluSpace Editing

Consider a large vision-language model (LVLM) that generates a token sequence $y = (y_1, \ldots, y_T)$ autoregressively conditioned on an image I and a prompt p. Object hallucination refers to cases where the generated text describes objects

that do not exist in the input image. Recent editing approaches attempt to mitigate hallucinations by identifying latent representation directions associated with hallucinated outputs and removing them from the model parameters. In particular, existing work [55] constructs a hallucination subspace by contrasting hallucinated and truthful responses under teacher-forcing evaluation. Formally, consider a dataset of paired responses,

$$\mathcal{D} = \{(x_i^+, x_i^-)\}_{i=1}^N, \tag{1}$$

where x_i^+ and x_i^- denote hallucinated and truthful responses respectively. Teacher-forcing is applied to obtain hidden representations at layer ℓ. Let $h_\ell(x)$ denote the pooled hidden representation of sequence x at layer ℓ. Stacking representations across samples yields matrices $X_\ell^+, X_\ell^- \in \mathbb{R}^{N\times D}$. The hallucination contrast matrix is defined as

$$E_\ell = X_\ell^+ - X_\ell^-. \tag{2}$$

Applying singular value decomposition (SVD), $E_\ell = U_\ell \Sigma_\ell V_\ell^\top$, the top-$k$ right singular vectors form the TF-based hallucination subspace as

$$V_\ell^{TF} = [v_1^{TF}, \ldots, v_{k_{TF}}^{TF}] \in \mathbb{R}^{D\times k_{\mathrm{TF}}}. \tag{3}$$

The model weights are then edited through null-space projection,

$$W_\ell^{edit} = (I - V_\ell^{TF}(V_\ell^{TF})^\top) W_\ell^{orig}, \tag{4}$$

which removes hallucination-related representation without additional training. However, TF-based HalluSpace is derived entirely from teacher-forced representations, whereas hallucinated outputs are produced during autoregressive decoding, where next-token decisions depend on local competitions among candidate tokens. As discussed in Sec. 1, this mismatch motivates incorporating AR-aware signals that reflect decoding-time dynamics.

3.2 AR-aware HalluSpace Construction

To better capture hallucination dynamics during generation, we construct hallucination directions directly from autoregressive decoding trajectories. We use the LURE dataset [63] as prior work. Given an image I and the question q = "Please describe this image in detail.", the LVLM generates a caption autoregressively as $y = (y_1, \ldots, y_T)$. Following the CHAIR [46] evaluation protocol, hallucinated objects in the generated caption are detected. Since CHAIR evaluates hallucination at the object level, we can identify the token position t where the hallucinated object first appears. Let $h_{\ell,j} \in \mathbb{R}^D$ denote the hidden representation at layer ℓ when generating token y_j. Instead of contrasting teacher-forced responses, we analyze local representation transitions around the hallucination position. Specifically, we compute sequential differences $\Delta h_{\ell,j} = h_{\ell,j} - h_{\ell,j-1}$. For each hallucination event at position t, we collect a short temporal window $\{\Delta h_{\ell,t}, \Delta h_{\ell,t-1}, \ldots, \Delta h_{\ell,t-w+1}\}$, which captures a short context of decoding-time

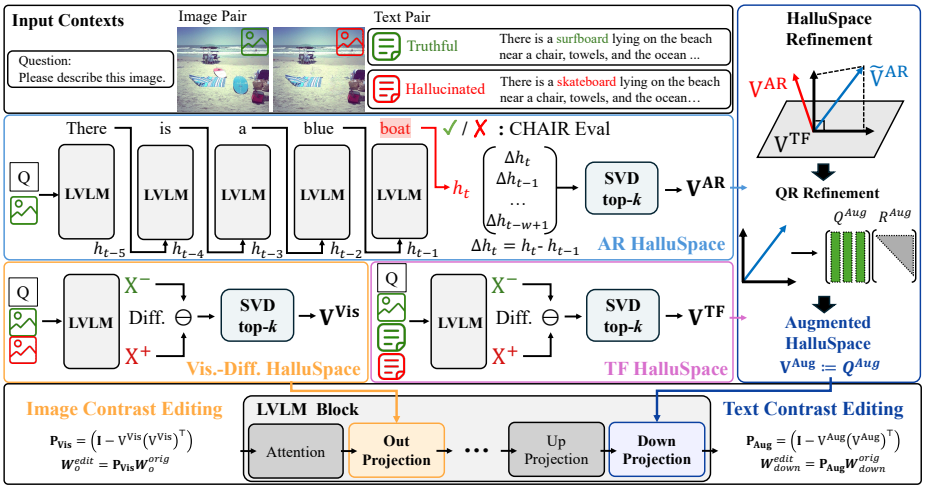

Fig. 2: Architecture overview of DARE. DARE constructs AR-aware, TF-based, and visual-difference HalluSpaces, refines the TF and AR directions into an Augmented HalluSpace, and applies module-specific null-space editing to mitigate hallucination.

state transitions leading into the hallucinated token. These vectors capture the representation dynamics immediately preceding hallucinated token generation. Collecting transitions from all hallucination samples forms the matrix

$$G_\ell = [\Delta h_\ell^{(1)}, \Delta h_\ell^{(2)}, \ldots, \Delta h_\ell^{(N_{\mathrm{AR}})}]^\top \in \mathbb{R}^{N_{\mathrm{AR}} \times D}, \tag{5}$$

where N_{AR} denotes the total number of collected transitions. Applying SVD to G_ℓ yields $G_\ell = U_\ell \Sigma_\ell V_\ell^\top$. Since transition vectors are stacked row-wise in G_ℓ, the right singular vectors in V_ℓ correspond to feature-space directions. The top-k_{AR} right singular vectors define the AR-aware hallucination subspace as

$$V_\ell^{\mathrm{AR}} = [v_1^{\mathrm{AR}}, \ldots, v_{k_{\mathrm{AR}}}^{\mathrm{AR}}] \in \mathbb{R}^{D \times k_{\mathrm{AR}}}. \tag{6}$$

Unlike TF-based HalluSpace, these directions capture autoregressive state transitions that occur immediately before hallucinated tokens are produced.

3.3 Augmented HalluSpace Editing

The TF-based HalluSpace and the AR-aware HalluSpace capture different aspects of hallucinated generation. To leverage these complementary signals, we construct an *Augmented HalluSpace* that integrates both TF-based and AR-based hallucination directions. Let V_ℓ^{TF} denote the TF HalluSpace obtained from teacher-forcing contrasts and V_ℓ^{AR} denote the AR HalluSpace derived from autoregressive generation trajectories. Before combining the two sets of directions, we residualize the AR basis vectors with respect to the TF subspace. This step ensures that AR directions capture information that is not already explained by

TF-based components and provides a non-redundant basis for null-space editing. Concretely, each AR vector v_j^{AR} is projected onto the orthogonal complement of the TF subspace as

$$\tilde{v}_j^{\mathrm{AR}} = \left(I - V_\ell^{\mathrm{TF}}(V_\ell^{\mathrm{TF}})^\top\right) v_j^{\mathrm{AR}}. \tag{7}$$

The resulting vectors are stacked to form the residualized AR basis $\tilde{V}_\ell^{\mathrm{AR}}$. We then concatenate the TF HalluSpace and the residualized AR directions and apply QR decomposition:

$$\left[V_\ell^{\mathrm{TF}}, \tilde{V}_\ell^{\mathrm{AR}}\right] = Q_\ell^{\mathrm{Aug}} R_\ell^{\mathrm{Aug}}. \tag{8}$$

Here, Q_ℓ^{Aug} contains the orthonormal basis vectors spanning the Augmented HalluSpace. For notational consistency with the HalluSpace bases, we denote this orthonormal basis as

$$V_\ell^{\mathrm{Aug}} := Q_\ell^{\mathrm{Aug}}. \tag{9}$$

Finally, DARE edits the FFN down-projection weights of the LVLM block using the corresponding null-space projection as follows:

$$P_\ell = I - V_\ell^{\mathrm{Aug}} {V_\ell^{\mathrm{Aug}}}^\top, \tag{10}$$

$$W_{\ell,\mathrm{down}}^{\mathrm{edit}} = P_\ell W_{\ell,\mathrm{down}}^{\mathrm{orig}}. \tag{11}$$

By applying this orthonormalized projection to the FFN down-projection layer, DARE suppresses hallucination-related representation directions while preserving the overall transformer architecture and inference pipeline.

3.4 Visual-Difference Editing

The analysis in Sec. 1 reveals that hallucination editing based solely on textual contrasts can unintentionally perturb visual components of the multimodal representation. As illustrated in Fig. 1(C), editing directions derived from hallucinated and truthful textual responses are accompanied by reduced attention to visual tokens across layers. This observation suggests that textual contrasts alone may not sufficiently capture the visual factors that contribute to hallucinated generations. To explicitly incorporate visual evidence into the editing process, we introduce *Visual-difference editing*, which constructs hallucination directions from controlled visual contrasts. We utilize the BEAF dataset [58], which provides paired images derived from MSCOCO [32] where specific objects are naturally removed from the original image. Each pair therefore forms a controlled visual counterfactual: one image contains the object while the other does not. Using MSCOCO annotations, we associate each pair with object-level ground-truth descriptions that reflect whether the removed object should appear in the caption. To isolate representation differences that arise purely from visual changes, hidden representations are extracted under teacher-forcing. Teacher-forcing ensures that

both images are processed under the same textual sequence, allowing representation differences to reflect variations in visual input rather than differences in generated language. This design is crucial because autoregressive decoding may produce different token sequences for the two images, which would entangle visual effects with generation dynamics. By aligning the textual context through teacher-forcing, we obtain a controlled representation contrast that isolates visual differences. Formally, we construct paired samples $\mathcal{D}_{vis} = \{(x_i^H, x_i^T)\}_{i=1}^N$, where x_i^H denotes the hallucinated condition and x_i^T denotes the truthful condition aligned with the visual content. Running teacher-forcing through the LVLM yields hidden representations at layer ℓ. After pooling over token positions, the resulting representations are stacked into matrices $X_\ell^H, X_\ell^T \in \mathbb{R}^{N\times D}$. We then define the visual contrast matrix as

$$E_\ell^{vis} = X_\ell^H - X_\ell^T, \tag{12}$$

which captures representation differences induced by controlled visual changes. We extract the dominant directions by decomposing E_ℓ^{vis} via SVD as

$$E_\ell^{vis} = U_\ell \Sigma_\ell (V_\ell^{vis})^\top. \tag{13}$$

The top-k_{vis} right singular vectors are selected to construct the visual-difference HalluSpace, denoted as $V_{\ell,k_{vis}}^{vis} = [v_1^{vis}, \ldots, v_{k_{vis}}^{vis}]$. Finally, the attention output projection weights are edited through null-space projection as

$$W_{\ell,o}^{edit} = \left(I - V_{\ell,k_{vis}}^{vis}(V_{\ell,k_{vis}}^{vis})^\top\right) W_{\ell,o}^{orig}. \tag{14}$$

Editing the attention output projection allows DARE to directly influence how visual features produced by the attention module are injected into the residual stream. This enables the editing process to suppress hallucination-related representations while preserving visually grounded information within the multimodal representation space.

4 Experiments

In this section, we evaluated our method against state-of-the-art approaches for mitigating object hallucination in LVLMs. We first describe the experimental setup in Sec. 4.1, followed by quantitative comparisons with existing methods in Sec. 4.2. Additional analyses and ablation studies are presented in Sec. 4.3.

4.1 Experimental Setup

Models. We conducted experiments on three widely used open-source LVLMs: LLaVA-1.5 [35], MiniGPT-4 [64], and mPLUG-Owl2 [57]. All experiments are performed using the official released checkpoints without additional fine-tuning. Unless otherwise specified, our primary analysis focuses on LLaVA-1.5.

Table 1: The **CHAIR** evaluation results of three baseline LVLMs with existing OH mitigation methods. The down-arrow (↓) indicates that lower $\mathbf{CHAIR}_S$ and $\mathbf{CHAIR}_I$ result in fewer object hallucinations, while the up-arrow (↑) indicates that higher **BLEU** result in better general generation quality. Bold and underlined indicates the best and the second best, respectively. We used 64 as the max token length.

Method	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	$\mathbf{CHAIR}_S \downarrow$	$\mathbf{CHAIR}_I \downarrow$	**BLEU** ↑	$\mathbf{CHAIR}_S \downarrow$	$\mathbf{CHAIR}_I \downarrow$	**BLEU** ↑	$\mathbf{CHAIR}_S \downarrow$	$\mathbf{CHAIR}_I \downarrow$	**BLEU** ↑
Greedy	24.32±2.92	8.91±0.46	15.16±0.30	35.91±3.15	14.61±0.09	14.43±0.46	24.96±1.43	9.16±1.04	14.74±0.62
Beam [16]	22.61±1.64	7.01±0.95	16.17±0.13	27.33±2.08	11.11±0.98	15.09±0.21	21.80±1.31	7.09±0.54	16.17±0.10
VCD [28]	21.15±1.25	7.19±0.26	14.91±0.46	30.60±1.95	11.06±1.11	14.94±0.22	20.56±2.40	7.01±0.34	15.46±0.19
OPERA [20]	17.61±1.16	<u>6.09±1.11</u>	15.21±0.49	29.64±0.18	10.91±0.76	15.42±0.52	20.19±0.15	7.16±0.13	15.31±0.64
SID [21]	17.95±1.93	6.64±0.11	15.04±0.24	27.16±2.06	10.34±0.46	15.83±0.27	19.66±1.43	7.13±0.18	15.94±0.16
VAF [59]	17.92±1.06	7.14±1.05	15.01±0.62	25.61±2.96	10.64±0.43	15.34±0.26	18.66±0.15	7.09±0.64	14.94±0.39
Nullu [55]	20.53±1.40	7.03±0.86	15.37±0.18	23.93±0.31	10.40±0.12	14.68±0.24	18.80±0.72	<u>6.86±0.30</u>	16.02±0.24
ONLY [53]	21.64±0.64	6.71±0.67	15.29±0.36	24.17±2.48	9.07±0.93	14.94±0.67	19.43±1.18	7.37±1.08	14.42±0.71
CMI-VLD [15]	<u>16.42±0.75</u>	6.59±0.81	15.40±0.86	<u>23.80±1.73</u>	<u>8.90±0.53</u>	15.38±0.33	<u>17.76±0.84</u>	6.95±0.94	15.11±0.34
DARE (Ours)	**14.93±1.68**	**6.07±0.82**	15.61±0.59	**22.64±1.49**	**8.73±0.96**	15.53±0.34	**16.81±0.43**	**6.57±0.84**	16.02±0.19

Baselines. We compared our method with several representative approaches for mitigating hallucinations in LVLMs. *Standard decoding baselines.* Greedy decoding and Beam Search [16] are included as standard generation baselines, reflecting model behavior without explicit hallucination mitigation. *Decoding-based methods.* VCD [28], OPERA [20], ONLY [53], and CMI-VLD [15] mitigate hallucination by modifying token selection during decoding or introducing inference-time constraints. *Representation intervention methods.* SID [21] perturbs intermediate vision representations and subtracts amplified hallucinated logits to encourage more factual predictions. VAF [59] reweights cross-attention scores in multimodal fusion layers to strengthen visual signals while suppressing system-prompt attention. *Model editing method.* Nullu [55] mitigates hallucinations by projecting model weights onto the null space of hallucination-related components, thereby removing hallucination-inducing directions without additional training.

Benchmarks. We evaluated hallucination mitigation performance using two widely adopted benchmarks for object hallucination: CHAIR [46] and POPE [31]. Both benchmarks use images from the MSCOCO validation set [32]. CHAIR measures the proportion of non-existent objects mentioned in generated captions, whereas POPE evaluates factual consistency through yes/no questions about object presence in images. To further assess general multimodal capabilities, we additionally reported results on MME [17], which evaluates perception and reasoning abilities across diverse multimodal tasks.

Implementation Details. All baseline implementations follow the official configurations provided in the HuggingFace Transformers repository. Additional implementation details are provided in the Supplementary Material.

4.2 Experimental Results

CHAIR Results. Table 1 shows that DARE consistently achieves the lowest hallucination scores across all evaluated LVLMs. On LLaVA-1.5, DARE reduces CHAIR_S from 24.32 to 14.93 and CHAIR_I from 8.91 to 6.07, showing substantial

improvements over both the base model and prior mitigation methods. Similar trends are observed on MiniGPT-4 and mPLUG-Owl2, indicating that the proposed editing strategy is not limited to a single backbone. Unlike TF-based editing methods such as Nullu [55], which derive hallucination directions from static teacher-forced contrasts, DARE incorporates AR-aware transitions that directly reflect generation-time dynamics. This allows DARE to better suppress hallucinated object mentions during actual autoregressive decoding while maintaining comparable generation quality, as reflected by the BLEU scores.

Table 2: The **POPE** with random sampling evaluation results on MSCOCO of three baseline LVLMs with existing OH mitigation methods. The up-arrow (↑) indicates that higher **Accuracy**, **Precision** and **F1 Score** result in better performance. Bold and underlined indicates the best and the second best, respectively. We used 64 as the max token number in this experiment. Results for random, popular and adversarial sampling, including recall metric are in the supplementary.

Method	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	Accuracy ↑	**Precision ↑**	**F1 Score ↑**	**Accuracy ↑**	**Precision ↑**	**F1 Score ↑**	**Accuracy ↑**	**Precision ↑**	**F1 Score ↑**
Greedy	79.53±1.56	82.64±0.84	82.47±0.71	72.84±1.41	70.79±1.15	69.68±0.93	75.71±0.90	77.68±1.35	80.71±0.48
Beam [16]	79.57±0.51	83.04±0.81	80.49±0.85	72.70±1.40	71.27±0.49	70.76±0.34	75.97±1.09	75.94±1.04	80.78±0.27
VCD [28]	81.61±0.29	82.90±0.95	82.79±1.11	75.62±0.46	75.06±0.77	78.17±0.85	78.48±1.30	79.34±0.91	83.62±0.55
OPERA [20]	87.83±1.31	88.34±0.84	88.70±0.90	75.46±0.33	75.60±0.18	79.30±0.39	78.03±0.59	81.90±0.47	84.03±1.64
SID [21]	87.90±1.32	86.64±0.47	88.03±0.33	76.41±0.67	79.67±0.08	79.45±0.48	78.26±0.80	82.39±1.10	84.63±0.79
VAF [59]	87.96±0.47	88.16±1.24	89.07±0.22	78.90±2.08	78.11±0.30	79.46±0.80	78.05±1.38	84.46±0.81	83.29±0.54
Nullu [55]	<u>88.73±1.43</u>	88.70±0.92	87.56±0.64	79.67±0.94	79.50±0.37	80.09±0.47	79.38±0.75	83.67±0.61	85.41±1.03
ONLY [53]	88.64±0.91	88.41±0.74	88.70±0.70	79.37±1.43	80.09±0.59	80.26±0.12	<u>81.66±0.90</u>	84.17±1.10	84.64±1.37
CMI-VLD [15]	88.01±0.17	<u>89.64±0.67</u>	<u>89.17±1.26</u>	<u>80.18±0.81</u>	<u>81.74±0.55</u>	<u>80.61±0.77</u>	81.59±0.21	<u>85.19±0.81</u>	<u>85.70±1.28</u>
DARE (Ours)	**89.74±0.54**	**91.40±0.71**	**90.04±0.37**	**80.96±0.49**	**81.97±0.40**	**81.67±0.75**	**82.94±1.07**	**85.36±0.18**	**85.78±0.48**

POPE Results. Table 2 further evaluates object-existence reasoning through yes/no questions. DARE consistently improves accuracy, precision, and F1 score across the evaluated LVLMs, demonstrating that its gains are not restricted to caption generation. Since POPE primarily measures whether models incorrectly affirm the presence of non-existent objects, these improvements indicate that DARE effectively reduces false-positive object predictions. Compared with TF-based editing, which captures representation differences under teacher forcing, DARE accounts for the model's actual autoregressive generation process. This leads to more reliable alignment between textual responses and visual evidence.

MME Results. Fig. 3 evaluates whether hallucination mitigation comes at the cost of general multimodal capability. DARE maintains competitive or improved MME performance compared with the base model and strong hallucination mitigation baselines. Since MME covers diverse perception and reasoning skills, including object existence, counting, position, color, OCR, and common-sense reasoning, these results suggest that DARE does not simply suppress object mentions or weaken visual understanding. Instead, DARE reduces hallucination while preserving the model's broader multimodal perception ability. This supports the design of combining TF/AR-based hallucination suppression with

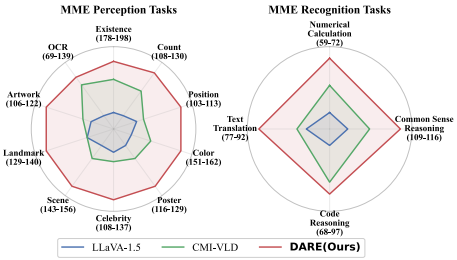

Fig. 3: Radar visualization of MME full-set evaluation results.

Table 3: Ablation study of the proposed DARE components.

Method	TF Editing	AR Editing	Visual Diff.	$\textbf{CHAIR}_S \downarrow$	$\textbf{CHAIR}_I \downarrow$
Base				24.32±2.92	8.91±0.46
(A)	✓			20.53±1.40	7.03±0.86
(B)		✓		17.61±1.64	6.94±1.06
(C)			✓	17.81±1.28	7.20±0.73
(D)	✓	✓		15.83±1.48	6.47±1.03
(E)		✓	✓	15.57±1.70	6.41±0.83
(F)	✓		✓	17.17±1.22	6.62±0.50
DARE (Ours)	✓	✓	✓	**14.93±1.68**	**6.07±0.82**

Table 4: Effects of editing layers $\{\ \ell\ \}$, rank k and window size for constructing AR-HalluSpace.

$\{\ \ell\ \}$	$\textbf{CHAIR}_S \downarrow$	$\textbf{CHAIR}_I \downarrow$	k	$\textbf{CHAIR}_S \downarrow$	$\textbf{CHAIR}_I \downarrow$	Window	$\textbf{CHAIR}_S \downarrow$	$\textbf{CHAIR}_I \downarrow$
16-31	**15.11**	**6.59**	1	18.44	7.80	1	16.92	6.87
24-31	16.87	6.93	**2**	**15.11**	**6.59**	2	16.15	6.82
0-15	17.70	7.22	4	17.36	6.90	**3**	**15.11**	**6.59**
0-7	17.37	7.18	8	17.09	6.83	4	15.84	6.63

visual-difference editing, which aims to mitigate hallucinated generations without disrupting visually grounded representations.

4.3 Ablation Studies and Further Analysis

Component Ablation. Table 3 presents an ablation study of the proposed DARE components. Each individual component improves over the base model: TF-based editing (A), AR-aware editing (B), and visual-difference editing (C) reduce both CHAIR_S and CHAIR_I, indicating that each provides a useful hallucination-mitigation signal. Among them, AR-aware editing achieves the strongest single-component performance, suggesting the importance of modeling hallucination directions that emerge during auto-regressive generation. Pairwise combinations further improve performance. TF+AR editing (D) outperforms either TF or AR alone, showing that teacher-forced and generation-aware hallucination signals are complementary. AR+visual-difference editing (E) gives the best pairwise result, while TF+visual-difference editing (F) also improves over TF alone, confirming that visual-difference cues help preserve image-grounded information during editing. Finally, combining all three components achieves the best overall performance, reducing CHAIR_S from 24.32 to 14.93 and CHAIR_I from 8.91 to 6.07. These results demonstrate that TF-based signals, AR generation dynamics, and visual-difference information provide complementary cues for effective hallucination mitigation.

Hyperparameter Sensitivity. Table 4 analyzes the sensitivity of AR-aware editing to the edited layers, HalluSpace rank, and temporal window size. Editing higher layers (16–31) performs best, suggesting that hallucination-related

Table 5: Generality evaluation on recent LVLMs with different model scales. DARE is evaluated on Qwen3-VL and Gemma-3 across 4B, 8B, and 12B variants, showing consistent improvements over the base model and Nullu on both CHAIR and POPE.

CHAIR Evaluation (C_S/C_I) ↓				
Method	Qwen3-VL-4B	Qwen3-VL-8B	Gemma-3-4B	Gemma-3-12B
Base	39.20 / 8.83	36.40 / 8.38	32.60 / 7.86	33.40 / 6.97
Nullu	37.80 / 8.44	34.80 / 7.48	30.50 / 7.01	32.56 / 6.29
DARE (Ours)	**36.75 / 7.80**	**33.17 / 6.51**	**29.67 / 6.78**	**30.91 / 6.08**
POPE Evaluation (Acc./F1) ↑				
Method	Qwen3-VL-4B	Qwen3-VL-8B	Gemma-3-4B	Gemma-3-12B
Base	89.99 / 89.40	88.88 / 88.22	84.30 / 84.53	84.33 / 85.28
Nullu	90.31 / 89.91	88.93 / 88.07	84.88 / 84.56	85.16 / 85.49
DARE (Ours)	**90.84 / 90.66**	**89.67 / 89.20**	**86.10 / 87.46**	**86.73 / 87.08**

directions are more effectively suppressed near the language generation stage. For the rank, $k = 2$ achieves the lowest CHAIR scores, while larger ranks slightly degrade performance, indicating that AR hallucination signals lie in a compact low-dimensional subspace. For the temporal window, a window size of 3 gives the best result, suggesting that hallucination-inducing transitions are concentrated within a short context before the hallucinated token.

Generality on Recent LVLMs. To further examine whether DARE generalizes beyond the LVLMs used in our main comparison, we additionally evaluate recent model families and scales, including Qwen3-VL-4B [5], Qwen3-VL-8B, Gemma-3-4B [51], and Gemma-3-12B. As shown in Table 5, DARE consistently improves over both the base model and Nullu on CHAIR and POPE across all evaluated backbones and model sizes. These results indicate that DARE is not limited to a specific LVLM family or parameter scale, and provides consistent gains for both caption-level hallucination and object-existence reasoning in recent LVLMs.

Throughput–Hallucination Trade-off. Fig. 4 compares generation throughput and CHAIR_S across hallucination mitigation methods. Decoding-based methods such as VCD and OPERA reduce hallucinations but require additional decoding operations, leading to lower throughput. In contrast, DARE achieves stronger hallucination mitigation while maintaining nearly the same generation speed as efficient editing-based methods. Compared with Nullu, DARE further reduces hallucination without introducing inference-time overhead, since the model is edited once through null-space projection and then decoded with the original inference pipeline. Thus, DARE provides a favorable balance between hallucination reduction and generation efficiency.

Visual Attention Drop after Editing. Fig. 5 analyzes how editing affects visual-token attention. Consistent with Fig. 1(C), TF-based editing causes a noticeable drop in visual attention, suggesting that text-derived hallucination directions can also perturb multimodal interactions. In contrast, DARE mitigates

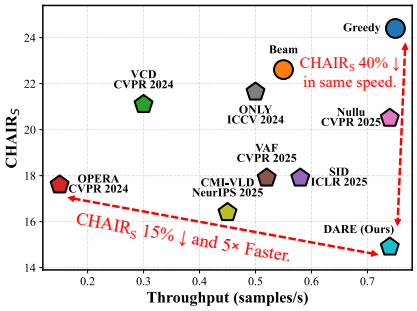

Fig. 4: Throughput–Hallucination Trade-off. Inference throughput and hallucination performance are measured on an NVIDIA RTX 4090.

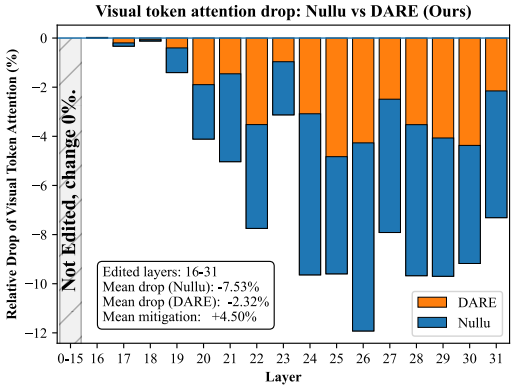

Fig. 5: Visual attention drop after proposed augmented HalluSpace and Visual-difference HalluSpace editing.

this drop: the average visual-token attention drop over layers 16–31 decreases from -7.53% for Nullu to -2.32% for DARE. The reduced attention drop indicates that DARE better preserves visual-token interactions during editing. This behavior is consistent with our design: TF and AR directions are integrated into the Augmented HalluSpace to suppress hallucination-related generation signals, while visual-difference editing complements them under visual-token disruption.

5 Conclusion

We presented DARE (Dual-path Auto-Regressive-aware Editing), a training-free weight editing framework for mitigating object hallucination in LVLMs. Motivated by our analyses of TF-based editing, we showed that teacher-forced representations can retain strong evidence for truthful tokens, yet TF-based editing often yields limited changes in the truthful–hallucinated decision margin under autoregressive decoding and can be accompanied by reduced attention to visual tokens. To address these gaps, DARE integrates dual-path contrasts—textual contrasts and image contrasts—with AR-aware signals derived from decoding trajectories. Concretely, DARE edits the FFN down-projection using an Augmented HalluSpace that combines TF and AR directions, and applies visual-difference editing on the attention output projection to better preserve visually grounded information flow. Extensive experiments on standard LVLM hallucination benchmarks demonstrated that DARE consistently reduces hallucinations while maintaining competitive multimodal performance and inference efficiency, since the editing is performed once without decoding-time overhead. Additional evaluations on recent LVLM families with different model scales further show that DARE is not limited to a specific backbone or parameter size. We hope DARE serves as a practical foundation for hallucination mitigation that goes beyond teacher-forced analysis and reflects autoregressive generation behavior.

Acknowledgements

This work was supported by Institute of Information and Communications Technology Planning and Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners), and supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT).

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Arora, K., El Asri, L., Bahuleyan, H., Cheung, J.C.K.: Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 700–710 (2022)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
6. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. Advances in neural information processing systems **28** (2015)
7. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
8. Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., Zhou, J.: Halc: Object hallucination reduction via adaptive focal-contrast decoding. arXiv preprint arXiv:2403.00425 (2024)
9. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
10. Choe, S.A., Park, K.H., Choi, J., Park, G.M.: Universal domain adaptation for semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4607–4617 (2025)
11. Choe, S.A., Shin, A.H., Park, K.H., Choi, J., Park, G.M.: Open-set domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23943–23953 (2024)
12. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J.R., He, P.: Dola: Decoding by contrasting layers improves factuality in large language models. In: ICLR (2023)

13. Dai, W., Liu, Z., Ji, Z., Su, D., Fung, P.: Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2136–2148 (2023)
14. Duan, J., et al.: Truthprint: Mitigating large vision-language models object hallucination via latent truthful-guided pre-intervention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
15. Fang, H., Zhou, C., Kong, J., Gao, K., Chen, B., Liang, T., Ma, G., Xia, S.T.: Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. NeurIPS (2025)
16. Freitag, M., Al-Onaizan, Y.: Beam search strategies for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation (2017)
17. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
18. Fu, Y., Xie, R., Sun, X., Kang, Z., Li, X.: Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 16563–16577 (2025)
19. Heo, J.W., Park, K., Park, G.M.: Detecting unknown objects via energy-based separation for open world object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27558–27567 (2026)
20. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
21. Huo, F., Xu, W., Zhang, Z., Wang, H., Chen, Z., Zhao, P.: Self-introspective decoding: Alleviating hallucinations for large vision-language models. In: Proceedings of the Thirteenth International Conference on Learning Representations (2025)
22. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: CVPR. pp. 27036–27046 (2024)
23. Jiang, N., Kachinthaya, A., Petryk, S., Gandelsman, Y.: Interpreting and editing vision-language representations to mitigate hallucinations. In: The Thirteenth International Conference on Learning Representations (2025)
24. Kim, J.M., Koepke, A., Schmid, C., Akata, Z.: Exposing and mitigating spurious correlations for cross-modal retrieval. In: CVPR. pp. 2584–2594 (2023)
25. Kim, M.J., Moon, J.Y., Sung, M., Park, G.M.: Open your model's eyes: Video and context-aware multimodal backchannel prediction. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2026)
26. Lee, T.Y., Park, S., Jeon, M., Hwang, H., Park, G.M.: Esc: Erasing space concept for knowledge deletion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5010–5019 (2025)
27. Lee, T.Y., Seo, J., Ko, J.H., Park, G.M.: Perturb a model, not an image: Towards robust privacy protection via anti-personalized diffusion models. Advances in Neural Information Processing Systems **38**, 92410–92442 (2026)
28. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR. pp. 13872–13882 (2024)

29. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
30. Li, X.L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., Lewis, M.: Contrastive decoding: Open-ended text generation as optimization. arXiv preprint arXiv:2210.15097 (2022)
31. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
33. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
34. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=J44HfH4JCg
35. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
36. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning **36** (2024)
37. Liu, S., Ye, H., Zou, J.: Reducing hallucinations in large vision-language models via latent space steering. In: ICLR (2024)
38. Liu, S., Zheng, K., Chen, W.: Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In: ECCV (2024)
39. Liu, Y., Ji, T., Sun, C., Wu, Y., Zhou, A.: Investigating and mitigating object hallucinations in pretrained vision-language (clip) models. arXiv preprint arXiv:2410.03176 (2024)
40. Lyu, X., Chen, B., Gao, L., Shen, H., Song, J.: Alleviating hallucinations in large vision-language models through hallucination-induced optimization. Advances in Neural Information Processing Systems **37**, 122811–122832 (2024)
41. Park, K.H., Choe, S.A., Park, G.M.: Sfuod: Source-free unknown object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3499–3508 (2025)
42. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 (2023)
43. Peng, S., Yang, S., Jiang, L., Tian, Z.: Mitigating object hallucinations via sentence-level early intervention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 635–646 (2025)
44. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. In: Proceedings of the Thirteenth International Conference on Learning Representations (2016)
45. Rawte, V., Sheth, A., Das, A.: A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922 (2023)
46. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. arXiv preprint arXiv:1809.02156 (2018)
47. Sarkar, P., Ebrahimi, S., Etemad, A., Beirami, A., Arik, S.O., Pfister, T.: Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In:

The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=yG1fW8igzP
48. Schmidt, F.: Generalization in generation: A closer look at exposure bias. In: Proceedings of the 3rd Workshop on Neural Generation and Translation. pp. 157–167 (2019)
49. Seo, J., Lee, S.H., Lee, T.Y., Moon, S., Park, G.M.: Generative unlearning for any identity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9151–9161 (2024)
50. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
51. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 technical report (2025), https://arxiv.org/abs/2503.19786
52. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
53. Wan, Z., Zhang, C., Yong, S., Ma, M.Q., Stepputtis, S., Morency, L.P., Ramanan, D., Sycara, K., Xie, Y.: Only: One-layer intervention sufficiently mitigates hallucinations

in large vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
54. Wang, C., Sennrich, R.: On exposure bias, hallucination and domain shift in neural machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3544–3552 (2020)
55. Yang, L., Zheng, Z., Chen, B., Zhao, Z., Lin, C., Shen, C.: Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
56. Yang, Z., Luo, X., Han, D., Xu, Y., Li, D.: Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
57. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In: CVPR. pp. 13040–13051 (2024)
58. Ye-Bin, M., Hyeon-Woo, N., Choi, W., Oh, T.H.: Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In: 18th European Conference on Computer Vision, ECCV 2024. pp. 232–248 (2025)
59. Yin, H., Si, G., Wang, Z.: Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
60. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4651–4659 (2016)
61. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024)
62. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839 (2023)
63. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. arXiv preprint arXiv:2310.00754 (2023)
64. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)