Look Closer! An Adversarial Parametric Editing Framework for Hallucination Mitigation in VLMs

Jiayu Hu[1], **Beibei Li**[1*], **Jiangwei Xia**[1], **Yanjun Qin**[2], **Bing Ji**[1], **Zhongshi He**[1]

[1]College of Computer Science, Chongqing University

[2]School of Computer Science and Technology, Research Center for Multimodal Information Perception and Intelligent Processing, Xinjiang University

{hujiayu, libeibeics, zshe}@cqu.edu.cn, xiajiangwei77@gmail.com, qinyanjun@xju.edu.cn

Abstract

While Vision-Language Models (VLMs) have garnered increasing attention in the AI community due to their promising practical applications, they exhibit persistent hallucination issues, generating outputs misaligned with visual inputs. Recent studies attribute these hallucinations to VLMs' over-reliance on linguistic priors and insufficient visual feature integration, proposing heuristic decoding calibration strategies to mitigate them. However, the non-trainable nature of these strategies inherently limits their optimization potential. To this end, we propose an adversarial parametric editing framework for Hallucination mitigation in VLMs, which follows an **A**ctivate-**L**ocate-**E**dit **A**dversarially paradigm. Specifically, we first construct an activation dataset that comprises grounded responses (positive samples attentively anchored in visual features) and hallucinatory responses (negative samples reflecting LLM prior bias and internal knowledge artifacts). Next, we identify critical hallucination-prone parameter clusters by analyzing differential hidden states of response pairs. Then, these clusters are fine-tuned using prompts injected with adversarial tuned prefixes that are optimized to maximize visual neglect, thereby forcing the model to prioritize visual evidence over inherent parametric biases. Evaluations on both generative and discriminative VLM tasks demonstrate the significant effectiveness of ALEAHallu in alleviating hallucinations.

Code — https://github.com/hujiayu1223/ALEAHallu

1 Introduction

In recent years, Vision-Language Models (VLMs) such as LLaVA (Liu et al. 2023b), MiniGPT-4 (Zhu et al. 2023), and BLIP-2 (Li et al. 2023a) have demonstrated remarkable capabilities across multimodal tasks including image-text generation, cross-modal retrieval, and visual question answering. Despite their impressive perceptual and generative capabilities, VLMs remain prone to hallucination issues (Liu et al. 2024b), where the model generates content inconsistent with the actual visual input on objects, attributes, and relationships, or even entirely unrelated. This issue greatly limits their reliability and performance in real-world applications.

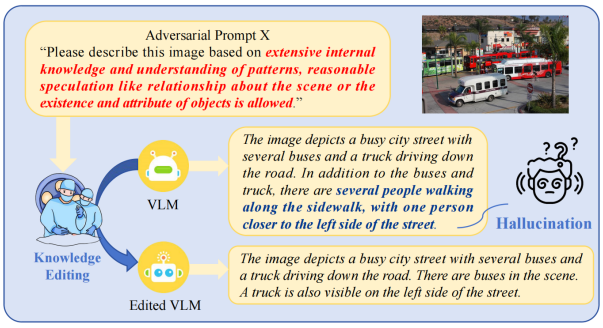

Figure 1: An example of ALEAHallu.

Hallucinations in VLMs stem from multiple factors (Liu et al. 2024b), including low-quality training data (Liu et al. 2023a), imperfect cross-modality alignment (Zhao et al. 2023), limitations in visual encoder architecture (Zhai et al. 2023), and biases inherent in Large Language Models (LLMs) (Ji et al. 2023). A critical contributor is the insufficient integration of visual features during generation (Wang et al. 2024a), coupled with an over-reliance on linguistic priors derived from text corpora (Liu et al. 2023a). This imbalance causes VLMs to prioritize linguistic fluency over factual accuracy, particularly when processing ambiguous visual inputs or out-of-distribution samples (Lee et al. 2023). In such cases, models often generate plausible yet factually inconsistent content that aligns with statistical language patterns rather than visual evidence (Wang et al. 2023), leading to hallucinations.

Recent work has explored decoding strategies that modulate generation dynamics to recalibrate VLMs' attention toward visual context while reducing dependence on LLM priors (Leng et al. 2024; Huang et al. 2024). While they demonstrate effectiveness, their heuristic, non-trainable nature, relying on manual decoding rules rather than learnable parameters, often leads to suboptimal performance and increased inference overhead, highlighting the need for optimizable solutions. Conversely, knowledge editing, which is a parameter-efficient technique that surgically modifies a pre-trained model's factual or behavioral knowledge at inference time without full retraining but by updating only a sparse set of its weights or activations, offers an efficient al-

*Corresponding author.

ternative. While knowledge editing has been applied to different LLM tasks (Wang et al. 2024b; Zhang, Yu, and Feng 2024; Chen et al. 2024b), their potential for VLM hallucination mitigation remains largely unexplored. A very recent work (Yang et al. 2025) employ knowledge editing in hallucination mitigation but relies on SVD decomposition of model parameters, suffering from high computation cost.

To this end, we propose an simple yet efficient ***A**ctivate–**L**ocate–**E**dit **A**dversarially* framework **ALEAHallu** that adversarially edits only those parameters responsible for vision hallucination. Specifically, we begin by curating an activation dataset of paired question-answer samples for each image: a *positive* response that is strictly grounded in the visual evidence, and a *negative* response that instead relies on parametric priors while ignoring the image. Then, processing these pairs in parallel, we compare the resulting hidden-state activations and statistically isolate the sparse parameter clusters whose activity is most correlated with hallucinatory outputs. Finally, we formulate an adversarial objective that forces the model to generate faithful answers under prompts with an adversarial prefix, encouraging the model to prioritize visual cues and thereby reduce hallucinations under standard prompts. Particularly, instead of hand-crafting such an adversarial prefix, we treat the prefix itself as a learnable matrix: keeping the model parameters fixed, we optimize the prefix tokens to maximize the probability of hallucinated responses.

Our contributions can be summarized as follows:

- **Dataset.** We construct a dataset specifically designed to localize hallucination-prone parameters, empowering systematic research into VLM hallucination mechanisms.
- **Methodology.** We propose an efficient adversarial parametric editing framework ALEAHallu for hallucination mitigation in VLMs, which follows an activate-locate-edit adversarially paradigm and edits only critical parameter clusters.
- **Experiments.** Extensive experiments across generative and discriminative tasks confirm that ALEAHallu significantly reduces hallucinations and robustly refocuses model attention on visual evidence.

2 Related Work

2.1 Visual-Language Models

To extend the capabilities of LLMs to vision-language tasks, pre-trained visual feature alignment and visual instruction fine-tuning are commonly used to help LLMs to comprehend the format of instruction input and generate diverse content in a more comprehensive way by integrating information from both text and images. Through these training process, more and more large visual language models (VLMs) emerged, including the series of CLIP (Liu et al. 2024b) and BLIP (Li et al. 2023a) well aligns the text features and image features and LLaVA (Liu et al. 2023b), LLaVA-NeXT(Liu et al. 2024a) and MiniGPT-4 (Zhu et al. 2023) could even allow users to interact with these intelligence with images and texts as prompts. Despite above advancements, specific challenges persist, especially the issue of object hallucination (Gunjal, Yin, and Bas 2024; Li et al. 2023b; Lovenia et al. 2023) being a prominent concern that affects the reliability and applicability of VLMs across domains. Consequently, we propose a novel method to mitigate hallucinations in vision-language models in our paper.

2.2 Hallucinations in VLMs

Hallucination in vision-language models (VLMs) appears as a discrepancy between generated text and image facts. Its origins span limited vision encoders, cross-modal misalignment, data bias, annotation irrelevance, and intrinsic LLM hallucinations (Hu et al. 2023; Liu et al. 2023a, 2024c). To alleviate this problem, recent research can be summarized into three lines of thought: at the training stage, by constructing positive and negative samples or introducing additional perceptual modalities to enhance the fine-grained nature and balance of visual instructions (Liu et al. 2023a; Jain, Yang, and Shi 2024), at the decoding stage, by leveraging contrasts among image–text (Zhu et al. 2024), inter-layer (Chuang et al. 2023), or original-vs.-perturbed (Leng et al. 2024) visual distributions to suppress the model's over-reliance on unimodal priors; other work attempts post-hoc correction with RLHF (Sun et al. 2023) or modified beam search (Huang et al. 2024). A very recent work (Wang et al. 2024b) eliminates VLM hallucinations via knowledge editing, yet they require SVD decomposition and incur heavy computational overhead. In a word, these methods either computationally expensive or rely on heuristic principles, making them impractical in real-world applications and prone to suboptimal outcomes.

2.3 Knowledge Editing

Knowledge editing techniques, which aim to efficiently adjust a model's behavior in specific domains while preserving overall performance, have garnered significant recent attention (Zhang et al. 2024). Current methods for editing large models are primarily categorized into three types (Zhang et al. 2024): Resorting to External Knowledge, Merging Knowledge into the Model, and Editing Intrinsic Knowledge. The first approach leverages prompting techniques, encouraging the model to reason using external knowledge provided within the input to generate improved responses (Zheng et al. 2023). Methods focusing on merging knowledge internally integrate representations of new knowledge directly into the model's parameters; acknowledging that feed-forward networks (FFNs) store significant knowledge (Geva et al. 2020, 2022; Chen et al. 2024a), several techniques specifically target modifications within these modules (Dong et al. 2022; Huang et al. 2023), while others directly replace specific internal representations (Hernandez, Li, and Andreas 2023). Editing Intrinsic Knowledge typically employs either Meta-Learning, utilizing an auxiliary hypernetwork to learn parameter adjustments based on new facts (De Cao, Aziz, and Titov 2021; Hase et al. 2023), or the Locate-and-Edit paradigm, which first identifies precise model locations storing specific knowledge before making targeted modifications (Meng et al. 2022a,b). Despite extensive research on knowledge editing for Large Language Models (LLMs), its application remains largely unexplored

in Vision-Language Models (VLMs), particularly for addressing hallucination mitigation.

3 Methodology

As illustrated in Figure 2, ALEAHallu comprises three core components: activation dataset construction, editing region localization, and adversarial parameter editing. Specifically, given a query (i.e., text prompt) and an image, we generate positive and negtive response pairs: a positive response that is visually grounded and hallucination-free, and a negative response that is over-reliance on LLM priors and containing hallucinations. These response pairs constitute the activation dataset and are utilized to pinpoint the critical hallucination-prone parameters that require editing. Subsequently, feeding the response pairs into the VLMs, we localize hallucination-prone parameters by analyzing the representational discrepancies between positive and negative samples in the latent space. Furthermore, we learn adversarial prompt prefix designed to overlook visual features and then edit hallucination-prone parameters to response correctly with adversarial prompt.

3.1 Activation Dataset Construction

Prompt design. To activate and localize hallucination-prone regions, we generate paired descriptions for the same image: one hallucination-free description that carefully attends to visual features, and one hallucinatory description containing erroneous objects, attributes, and relationships due to the neglect of visual information. Specifically, for each image v, we design a visually grounded prompt x^+ as shown in Prompt 1, focusing on visual features and eliciting a less hallucinated response y^+, forming the positive sample $[v, x^+, y^+]$. Conversely, for negative samples, we use an adversarial prompt x^- as shown in Prompt 2, which focuses more on knowledge priors and is designed to trigger hallucinations, and concatenate it with the corresponding hallucinated response y^- to form $[v, x^-, y^-]$. The detailed prompts are as follows.

GPT-4o assisted evaluation. To ensure that the positive sample does not contain hallucinations, we need to help evaluate the level of hallucination in the response. In this paper, we use GPT-4o as our evaluation tool. The detailed prompt is as Prompt 3.

Prompt 1: Focus on visual feature

Please describe the given image based on specific visible information and avoid any imagination that is not in the image.

Prompt 2: Focus on language priors

Please describe the given image based on extensive internal knowledge and understanding of patterns, reasonable speculation like relationship about the scene or the existence and attribute of objects is allowed.

Prompt 3: GPT-4o evaluation prompt

You are required to score the hallucination degree of description of a given image. Please score 0∼2 based on the principle of scoring.

Principle of scoring:
• the description is *precise and everything in the description is all visible in the image*: 0.
• there are *wrong relationship and attribute of objects* but the *objects in description are all in the image*: 1.
• there are *objects that do not exist in image but appear in the given description*: 2.
Please output the scores for the description: [the generated image caption].
Do not generate other sentence after the score.

Output format:
Degree: <Scores of the description>

Samples with a score of 0 are hallucination-free and treated as positive samples. Samples scored 1 or 2 exhibit varying degrees of hallucinations and are considered negative samples. For an image, if the descriptions generated based on Prompt 1 and Prompt 2 correspond to hallucination-free and hallucinated outputs respectively, they form a positive-negative sample pair. In practice, we utilized Prompt 1 and Prompt 2 to generate descriptions for 4,000 images selected from MSCOCO. After verification by GPT-4o, 2,091 valid positive-negative sample pairs were obtained.

3.2 Editing Region Localization

We adopt a simple strategy that leverages the representation of an entire response to localize the editing region. As we know, a Transformer consists of multiple decoder layers. During decoding, the input query prompt and the corresponding image are first encoded into embeddings and then forwarded through the layers. Each layer comprises a multi-head self-attention module and multi-layer perceptrons.

To identify hallucination-prone regions, we feed the positive–negative response pairs constructed in the previous section into the VLM and compute the layer-wise Euclidean distance between their hidden representations. The layer exhibiting the largest distance is regarded as the hallucination-prone layer, denoted ℓ_{hallu}:

$$\ell_{\text{hallu}} = \underset{\ell \in \{1,\dots,L\}}{\operatorname{argmax}} \left\| \mathbf{h}_\ell^{(+)} - \mathbf{h}_\ell^{(-)} \right\|_2, \tag{1}$$

where $\mathbf{h}_\ell^{(+)}, \mathbf{h}_\ell^{(-)}$ are the hidden states of the positive and negative responses at layer ℓ, respectively.

Typically, the MLP in each decoder layer is a two-layer feed-forward network. Following prior work (Wang et al. 2024b), we treat the weight matrix of the second layer within the MLP in ℓ_{hallu} -th layer as the hallucination-prone parameters, since this layer plays a pivotal role in knowledge dissemination during forward propagation and serves as the target of our subsequent editing.

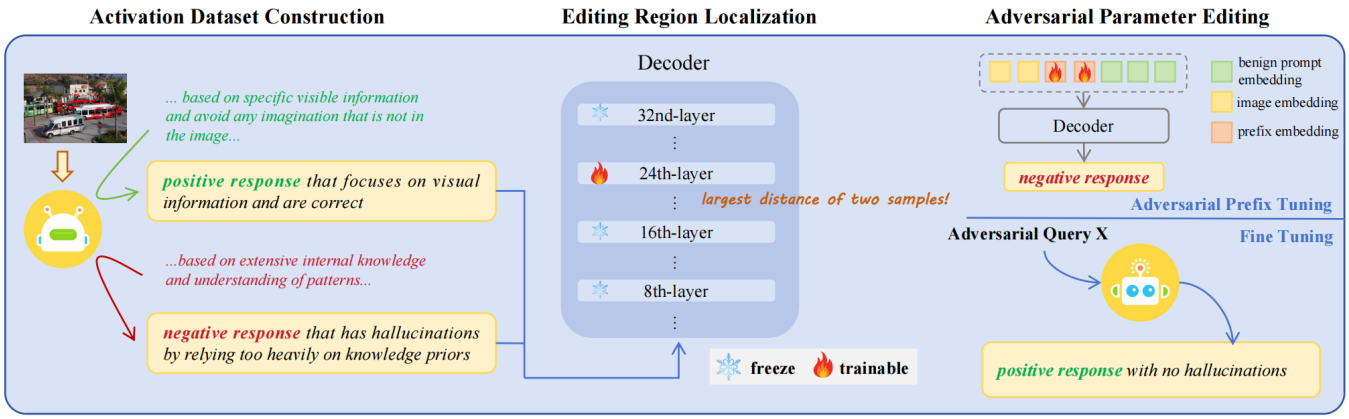

Figure 2: ALEAHallu consists of three main components: activation dataset construction, editing region localization and adversarial parameter editing.

3.3 Adversarial Parameter Editing

To edit the identified hallucination-prone region, we formulate an optimization objective that trains the model to remain robust, i.e., to generate hallucination-free content, even under adversarial input conditions. Such adversarial inputs include prompts that either encourage the model to ignore visual information or distract the model from attending to the correct visual cues. Our insight is that, *to produce visually consistent outputs under such perturbations, the model is forced to attend more accurately to the visual evidence, thereby mitigating hallucinations under benign prompts, which are much easier than adversarial prompts*. Intuitively, if a person can answer questions correctly under strong distractions, she/he will certainly perform even better in distraction-free settings.

Let the benign prompt be x, for example, in image-captioning, x could be "*Please describe the following image*". We construct an adversarial prompt prefix, denoted q, e.g., "*Please rely more on the LLM prior rather than the image content*". Concatenating q and x yields the adversarial prompt $[q; x]$, which is supplied to the VLM together with the image v. We minimize the negative log-likelihood of the ground-truth non-hallucinated response y^+ under this adversarial prompt:

$$\min_{\mathbf{W}_t} \mathcal{L}_e = -\log P_{\mathbf{W}_t}\left(y^+ \mid v, q, x\right), \tag{2}$$

where $\mathbf{W}_t$ denotes the hallucination-prone parameters at training step t. This objective encourages the model to generate accurate responses despite misleading instructions, progressively reducing hallucinations under benign scenarios.

Meanwhile, to prevent knowledge editing from distorting the original output distribution, we incorporate a KL-divergence regularization term as Equation 3, which constrains the edited parameters $\mathbf{W}_t$ so that the model retains its ability to respond normally to benign user queries and to produce fluent sentences. This ensures that localized parameter updates do not degrade general performance.

$$\min_{\mathbf{W}_t} \mathcal{L}_c = \mathrm{KL}\left(P_{\mathbf{W}^t}\left(\cdot \mid [v, x]\right) \| P_{\mathbf{W}_0}\left(\cdot \mid [v, x]\right)\right), \tag{3}$$

where $\mathbf{W}_0$ denotes the original model parameters.

Adversarial Prefix Tuning. Manually crafting adversarial prompt prefixes is heuristics, labor-intensive and expertise-dependent. We therefore employ prompt tuning to automatically optimize an adversarial prefix. Our core idea involves: keeping the VLM parameters frozen, prepending a learnable prefix to the input prompt, and optimizing this prefix to induce hallucinatory responses resulting from the model's neglect of visual information. An adversarial prefix optimized under this objective is supposed to guide the prompt to disregard visual cues.

We parameterize the prefix as a continuous matrix $\mathbf{E}_x \in \mathbb{R}^{r \times d}$, where r is the number of prefix tokens and d is the word-embedding dimension of the VLM. This matrix is concatenated with the image embedding $\mathbf{V}$ and the benign prompt embedding $\mathbf{E}_p$, forming the input $[\mathbf{V}; \mathbf{E}_q; \mathbf{E}_x]$ to the VLM. To train the prefix, we freeze all the parameters of VLM and maximize the likelihood of the hallucinated response y^- from our dataset as follows:

$$\min_{\mathbf{E}_q} \mathcal{L}_q = -\log P_{\mathbf{W}_0}\left(y^- \mid [\mathbf{V}, \mathbf{E}_q, \mathbf{E}_x]\right). \tag{4}$$

Finally, when the edited model operates under benign prompts (significantly easier than adversarial prompts), the model will pay more attention to visual tokens and the hallucinations are mitigated naturally.

3.4 Training and Inference

Our training proceeds in two stages. *Stage 1*: We perform prefix tuning by optimizing Equation 4 to obtain the adversarial prefix $\mathbf{E}_q$. *Stage 2*: We jointly optimize the parameter editing objective $\mathcal{L}_e$ together with the KL regularization term $\mathcal{L}_c$ as follows:

$$\min_{\mathbf{W}_t} \mathcal{L} = \min_{\mathbf{W}_t} (\lambda \mathcal{L}_e + \mathcal{L}_c), \tag{5}$$

where λ is a scalar weight.

At inference time, we simply feed the image and the benign prompt into the VLM and generate output.Note that during the training phase we freeze every parameter except

Methods	LLaVa-1.5				LLaVa-NeXT				MiniGPT-4			
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Recall↑	Len	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Recall↑	Len	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Recall↑	Len
Regular	56.4	17.2	69.5	106.2	40.4	12.2	60.2	175.9	56.4	17.2	69.5	106.2
BS	50.6	13.5	**78.8**	97.6	35.2	8.5	62.8	176.7	**34.8**	**10.0**	61.1	80.6
OPERA	51.3	13.6	74.2	93.6	35.2	8.5	62.8	176.7	**34.8**	**10.0**	61.1	80.6
VCD	57.4	16.1	76.7	103.1	34.6	9.6	60.7	175.6	57.4	16.1	76.7	103.1
Nullu	46.8	14.9	67.9	95.0	35.5	10.0	61.3	173.5	48.4	13.5	56.8	105.2
ALEAHallu	**39.3**	**12.1**	74.2	98.1	**34.4**	**7.6**	**63.0**	185.6	40.2	13.0	**78.0**	98.1

Table 1: Results on image caption task using different base models, the values in bold and underlined are the best and second best results in each row.

those in the hallucination-prone region, and our method is identical to the original model at inference and incurs zero additional decoding overhead. Hence, it is highly efficient.

4 Experiment

4.1 Settings

Datasets and Metrics

Image Caption Tasks. We conduct experiments of image caption tasks on the MSCOCO dataset (Lin et al. 2014) by querying the same prompt 'Please describe this image in detail', which contains over 300,000 images and 80 object categories with annotations. We use CHAIR(Rohrbach et al. 2018) to to quantify the degree of object hallucination in image captioning by determining the proportion of objects mentioned in the generated description but absent from the ground-truth set. It provides two metrics: $CHAIR_s$ and $CHAIR_i$, to measure hallucinations at the sentence and image levels, respectively. We also evaluate the responses using Recall and description length.

VQA Tasks. We choose POPE evaluation, which designs binary questions about object presence in images. They all include three sampling settings: random, popular, and adversarial. We select the MSCOCO, A-OKVQA (Schwenk et al. 2022), and GQA (Hudson and Manning 2019) for our POPE benchmark and adopt Accuracy, Precision, Recall, and F1 score as our evaluation metrics. The three sources involve 500 images from each dataset under each sampling setting, with six questions formulated per image. Since the negative samples can be easily constructed by flipping the binary "yes/no" answers from the corresponding positive samples in this setting, we select 2,000 pairs and use 80% of them as training data, 20% of them as test data.

For overall performance assessment, we introduce The Multi-modal Large Language Model Evaluation (MME) benchmark (Fu et al. 2023), which contains ten perception-related and four cognition-related tasks, encompassing both object-level and attribute-level hallucinations. We select perception-related tasks in our experiments to test the edited models.

Compared Methods We compare ALEAHallu with five baselines: 1) **Direct Sampling**: The next token is directly sampled from the post-softmax distribution. 2) **Beam Search (short as BS)** (Wu et al. 2016): It keeps the top-k most likely sequences at each step instead of just the best one and expands all current beams by possible next tokens and retains the top-k overall. 3) **OPERA** (Huang et al. 2024): It utilize a penalty term on the model logits during the beam-search decoding, along with a rollback strategy that retrospects the presence of summary tokens in the previously generated tokens, and re-allocate the token selection if necessary. 4) **VCD** (Leng et al. 2024): It calibrates the model's outputs by contrasting output distributions derived from original and distorted visual inputs. 5) **Nullu** (Yang et al. 2025): It identifies a subspace by extracting the hallucinated embeddings features and removing the truthful representations. Input features will be projected into the Null space of it by orthogonalizing the model weights.

Implementation Details. We choose 500 positive and negative sample pairs to locate the editing regions and train our model on 8 NVIDIA GeForce RTX 3090 Ti GPUs, setting the batch size, epochs to 10, 5 respectively, each epoch runs about 1045 seconds. We adopt LLaVa-1.5, LLaVa-NeXT and MiniGPT-4 as base model in our experiments and set LLaVA-1.5 as default one. We set the learning rate and weight decay to 2×10^{-5}, max new tokens is set to 512, λ is set to 0.1. For Beam Search and OPERA decoding, we set the number of beams to 2, VCD and Nullu are under their default configurations. The length of our learnable prefix is 5 tokens. 1,500 samples are used to optimize the prefix.

4.2 Experimental Results

Results on Image Caption Tasks. We provide the experimental results of image caption tasks on three models: LLaVa-1.5, LLaVa-NeXT and MiniGPT-4 as shown in Table 1. ALEAHallu maintains the lowest or second-lowest hallucination rates while achieving the highest recall among all methods.

For the metrics $CHAIR_s$ and $CHAIR_i$ on LLaVa-1.5, the two knowledge editing methods, i.e., Nullu and ALEAHallu, outperform all other compared method based on decoding strategy, indicating that knowledge editing has advantages in reducing object hallucinations and our method could achieve the better performance. Regarding Recall, beam search achieves the highest score of 78.8 on LLaVa-1.5, suggesting that it can generate the most ground-truth objects from the images. Our method achieves a Recall of 74.2, slightly lower but still outperforms the other baselines. This may be attributed to the broader search space provided by beam search, allowing it to capture a wider range of ob-

Datasets	Methods	Random	Popular	Adversarial
MSCOCO	ALEAHallu	**0.9072**	**0.8952**	0.8016
	Regular	0.8452	0.8288	0.8076
	Beam Search	0.8752	0.8692	**0.8528**
	OPERA	0.8744	0.8684	0.8520
	VCD	0.8652	0.8464	0.8252
	Nullu	0.8499	0.8200	0.7900
A-OKVQA	ALEAHallu	0.8832	**0.8724**	**0.8020**
	Regular	0.8516	0.8088	0.7388
	Beam Search	0.8852	0.8412	0.7812
	OPERA	0.8868	0.8420	0.7856
	VCD	0.8696	0.8124	0.7488
	Nullu	**0.8957**	0.8600	0.7800

Table 2: Results of Accuracy on VQA tasks, the values in bold and underlined are the best and second best results in each column.

jects across different generated sequences.

For LLaVa-NeXT, hallucinations decrease across all methods on this stronger model and our method still delivers the best performance——the lowest CHAIR_s, CHAIR_i and the highest Recall. Meanwhile, the sequence lengths generated by ALEAHallu are comparable to those of other approaches, indicating that the decreasing in CHAIR_s and CHAIR_i are not simply due to longer descriptions, thereby ensuring a fair comparison.

As for MiniGPT-4, although certain baselines (e.g., BS and OPERA) obtain slightly lower CHAIR_i than ours, their recall drops substantially (e.g., 10 vs. 13 for CHAIR_i, but recall only 61.1 vs. 78 for ALEAHallu). We could also find that OPERA has limited ability in this model as well as MiniGPT-4 and obtain the same results as Beam Search's.

Results on VQA Tasks. According to the results in Table 2, ALEAHallu generally outperforms other approaches across different settings, particularly under the popular scenario, demonstrating a clear advantage in reducing object hallucinations. Our approach is also competitive in rest cases, highlighting its robustness across diverse query distributions. While ALEAHallu delivers the best overall performance, Beam Search and OPERA decoding also show their advantages in certain settings, indicating that decoding strategies still play a role in mitigating hallucinations to some extent.

Results on MME. We present the results of ten perception-related tasks in Figure 3 and use LLaVA-1.5 as a representative. Our methods achieves better performance on most tasks, especially for *color*, *position*,*count* and *existence* task which evaluates the performance in reducing object-level and attribute-level hallucinations. The other tasks results also shows that our method leads to an enhancement of the general capability in perception-related tasks.

These results suggest that ALEAHallu not only effectively aligns generated content with visual features but also generalizes well across different hallucination scenarios, validating its effectiveness in addressing the core challenges of hallucination in VLMs.

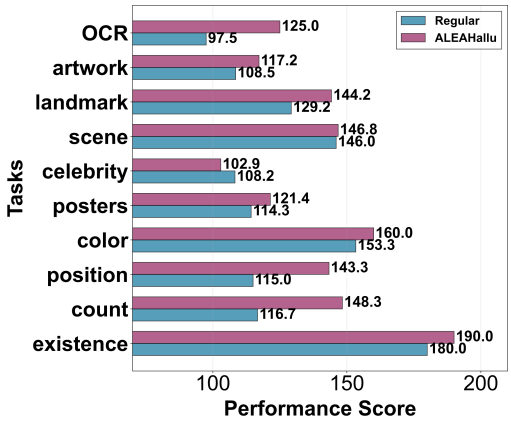

Figure 3: Results on MME.

Methods	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Recall ↑	Len
ALEAHallu	**39.3**	**12.1**	**74.2**	**92.2**
Regular	56.4	17.2	69.5	106.2
w/o Tune	45.8	12.9	78.5	93.7
w/o Location	55.4	16.1	78.4	100.5
w/o Editing	52.0	15.3	79.3	97.0
w/o Constraint	48.2	13.6	77.1	100.5
w/o Prefix Tuning	48.4	13.0	78.0	98.1

Table 3: Ablation study.

4.3 Ablation Study

We assess the effectiveness of each component in ALEAHallu in ablation study. We evaluate the impact of parameter tuning, where only the prompt that focuses more on the visual feature is used without fine-tuning the model (w/o Tune). Additionally, we investigate the effect of locating the hallucination region (w/o Location) by randomly selecting one layer for editing instead of using our proposed identification method. To assess the contribution of the loss functions, experiments are conducted by individually removing the knowledge editing loss ($\mathcal{L}_e$), the knowledge constraint loss ($\mathcal{L}_c$) and prompt learning, denoted as w/o Editing, w/o Constraint and w/o Prefix Learning, respectively. The results are shown in Table 3.

Editing and Constraint losses play important roles in ALEAHallu. Omitting parameter tuning results in significantly higher CHAIR_s and CHAIR_i scores, indicating increased object hallucination. This demonstrates the necessity of targeted fine-tuning. Additionally, removing either $\mathcal{L}_e$ or $\mathcal{L}_c$ results in reduced performance, they still have a better performance than regular method but worse than ALEAHallu, even though w/o Editing obtains higher Recall, that is because without the restraint of $\mathcal{L}_c$, it could not only generate more possibly existed object but also more hallucinated objects, highlighting the complementary roles of both losses in effectively guiding the training process.

Locating the correct layers for editing is essential. Editing a randomly selected layer instead of our localization ap-

Dataset	Model	Accuracy	Precision	Recall	F1 score
MSCOCO	$E_{\text{A-OKVQA}}$	**0.9060**	**0.9042**	**0.9108**	**0.9074**
	E_{GQA}	0.8948	0.8942	0.8263	0.8886
	Regular	0.8452	0.8488	0.8316	0.8502
A-OKVQA	E_{MSCOCO}	0.8732	0.8234	**0.9568**	0.8850
	E_{GQA}	**0.9100**	**0.9022**	0.9239	**0.9128**
	Regular	0.8516	0.8256	0.8820	0.8701
GQA	E_{MSCOCO}	**0.8736**	**0.8212**	0.9616	**0.8858**
	$E_{\text{A-OKVQA}}$	0.8516	0.7862	**0.9742**	0.8701
	Regular	0.8324	0.7900	0.8964	0.8632

Table 4: The generalization ability across datasets.

proach leads to a noticeable performance drop, confirming the importance of accurate hallucination region identification. Specifically, the higher layers which contains most semantic information are selected more in our experiments, randomly choose one layer in shallow, middle layers would not well guide the representation to a more image focus direction.

Prefix Tuning helps enhance the capability of ALEAHallu. Using manually designed prompt rather than learnable prompt would reduce performance in generating non-hallucinatory responses, however, it could capture more existent objects information in images since its Recall is higher than ALEAHallu.

4.4 Generalization Analysis

We conduct cross-dataset experiments on the POPE benchmark to investigate whether edits based on one dataset (e.g., MSCOCO) can generalize to other datasets (e.g., GQA). Specifically, we train three edited models, i.e., E_{MSCOCO}, $E_{\text{A-OKVQA}}$, and E_{GQA}, each fine-tuned on hallucination-prone samples from the respective dataset. The results are shown in Table 4. It demonstrates that our editing model, when applied to one dataset, could also enhance the ability to mitigate hallucinations between datasets, which shows the superiority of ALEAHallu.

4.5 Visual Attention Analysis

This paper aims to enhance visual attention in VLM through knowledge editing, thereby mitigating hallucinated outputs. Experimental results on a test set of 500 samples show that the average proportion of attention allocated to the image by the VLM increased from 29.71% to 33.72%.

To provide a more intuitive understanding of the change in attention before and after knowledge editing, we visualize the attention scores between the last input token and each image token. In these visualizations, brighter regions indicate higher attention weights. As illustrated in Figure 4, the high-attention regions become both more prominent and more widespread after editing, suggesting that the VLM pays significantly more attention to the image content following the knowledge edit.

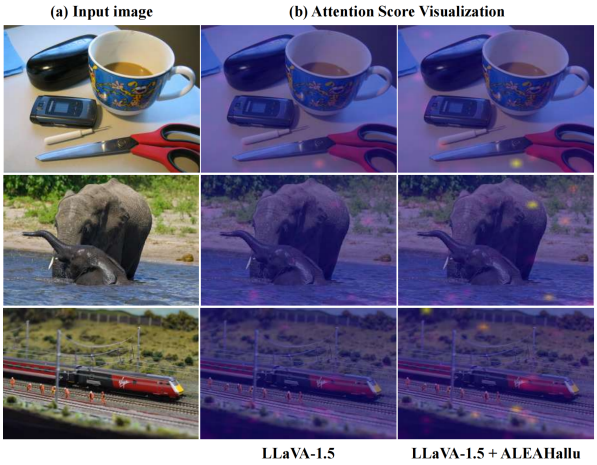

Figure 4: Visualization of Visual Attention Scores.

5 Conclusion

In this paper, we address the persistent hallucination problem in VLMs and propose **ALEAHallu**, an adversarial parametric editing framework for hallucination mitigation. By following an *activate-locate-edit adversarially paradigm*, ALEAHallu efficiently reduces hallucinations through edits to only a minimal parameter subset, while imposing no additional inference overhead. Extensive experiments demonstrate that ALEAHallu outperforms existing approaches while significantly enhancing VLMs' attentional focus on visual evidence during generation, thereby validating its effectiveness. For future work, we plan to integrate curriculum learning to further explore the performance ceiling of knowledge editing for hallucination suppression in VLMs.

6 Broader Impacts

Our work propose a hallucinations mitigating method for visual language models, which has a broader impact on enhancing trust in AI systems by reducing misinformation, making VLMs safer and more reliable in practical application. particularly in high-stakes domains like healthcare, education, and media. In education, hallucination mitigation may enhance learning tools by grounding visual-textual explanations in factual content. Nevertheless, over-mitigation of hallucinations could introduce trade-offs. In creative domains, suppressing imaginative synthesis may limit expressive potential. The societal impact of hallucination mitigation ultimately depends on balancing factual alignment with transparency, interpretability, and the preservation of human agency in evaluating AI outputs.

Acknowledgements

This work was supported by the Chongqing Science and Technology Bureau (CSTB2022TTAD-KPX0180); "Tianchi Yingcai" Introduction Program;the National Natural Science Foundation of China Nos.62306164; Basic Research Project of the Autonomous Region's Universities' Basic Research Operating Funds (XJEDU2025J001); Open Research Fund Program of Beijing National Research

Center for Information Science and Technology; Key Research and Development Project of the Autonomous Region(2024B03028).

References

Chen, Y.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2024a. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17817–17825.

Chen, Z.; Sun, X.; Jiao, X.; Lian, F.; Kang, Z.; Wang, D.; and Xu, C. 2024b. Truth forest: Toward multi-scale truthfulness in large language models through intervention without tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20967–20974.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; and He, P. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

De Cao, N.; Aziz, W.; and Titov, I. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*.

Dong, Q.; Dai, D.; Song, Y.; Xu, J.; Sui, Z.; and Li, L. 2022. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Geva, M.; Caciularu, A.; Wang, K. R.; and Goldberg, Y. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*.

Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Gunjal, A.; Yin, J.; and Bas, E. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18135–18143.

Hase, P.; Diab, M.; Celikyilmaz, A.; Li, X.; Kozareva, Z.; Stoyanov, V.; Bansal, M.; and Iyer, S. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2714–2731.

Hernandez, E.; Li, B. Z.; and Andreas, J. 2023. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.

Hu, H.; Zhang, J.; Zhao, M.; and Sun, Z. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13418–13427.

Huang, Z.; Shen, Y.; Zhang, X.; Zhou, J.; Rong, W.; and Xiong, Z. 2023. Transformer-patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.

Jain, J.; Yang, J.; and Shi, H. 2024. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27992–28002.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

Lee, S.; Park, S. H.; Jo, Y.; and Seo, M. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

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. In *International conference on machine learning*, 19730–19742. PMLR.

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 preprint arXiv:2305.10355*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Liu, Y.; Ji, T.; Sun, C.; Wu, Y.; and Zhou, A. 2024c. Investigating and Mitigating Object Hallucinations in Pretrained Vision-Language (CLIP) Models. arXiv:2410.03176.

Lovenia, H.; Dai, W.; Cahyawijaya, S.; Ji, Z.; and Fung, P. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35: 17359–17372.

Meng, K.; Sharma, A. S.; Andonian, A.; Belinkov, Y.; and Bau, D. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, 146–162. Springer.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Wang, B.; Wu, F.; Han, X.; Peng, J.; Zhong, H.; Zhang, P.; Dong, X.; Li, W.; Li, W.; Wang, J.; et al. 2024a. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5309–5317.

Wang, J.; Zhou, Y.; Xu, G.; Shi, P.; Zhao, C.; Xu, H.; Ye, Q.; Yan, M.; Zhang, J.; Zhu, J.; et al. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Wang, M.; Zhang, N.; Xu, Z.; Xi, Z.; Deng, S.; Yao, Y.; Zhang, Q.; Yang, L.; Wang, J.; and Chen, H. 2024b. Detoxifying Large Language Models via Knowledge Editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3093–3118.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang, L.; Zheng, Z.; Chen, B.; Zhao, Z.; Lin, C.; and Shen, C. 2025. Nullu: Mitigating Object Hallucinations in Large Vision-Language Models via HalluSpace Projection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhai, B.; Yang, S.; Zhao, X.; Xu, C.; Shen, S.; Zhao, D.; Keutzer, K.; Li, M.; Yan, T.; and Fan, X. 2023. HallE-Switch: Rethinking and Controlling Object Existence Hallucinations in Large Vision-Language Models for Detailed Caption.

Zhang, N.; Yao, Y.; Tian, B.; Wang, P.; Deng, S.; Wang, M.; Xi, Z.; Mao, S.; Zhang, J.; Ni, Y.; et al. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Zhang, S.; Yu, T.; and Feng, Y. 2024. TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8908–8949.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Zheng, C.; Li, L.; Dong, Q.; Fan, Y.; Wu, Z.; Xu, J.; and Chang, B. 2023. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint 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 preprint arXiv:2402.18476*.