

Through the Magnifying Glass: Adaptive Perception Magnification for Hallucination-Free VLM Decoding

Shunqi Mao Chaoyi Zhang Weidong Cai

School of Computer Science, The University of Sydney, Australia

{shunqi.mao, chaoyi.zhang, tom.cai}@sydney.edu.au

Abstract

Existing vision-language models (VLMs) often suffer from visual hallucination, where the generated responses contain inaccuracies that are not grounded in the visual input. Efforts to address this issue without model finetuning primarily mitigate hallucination by contrastively reducing language biases or amplifying the weights of visual embedding during decoding. However, these approaches remain limited in their ability to capture fine-grained visual details. In this work, we propose the Perception Magnifier (PM), a novel visual decoding method that iteratively isolates relevant visual tokens based on attention and magnifies the corresponding regions, spurring the model to concentrate on fine-grained visual details during decoding. By magnifying critical regions while preserving the structural and contextual information at each decoding step, PM allows the VLM to enhance its scrutiny of the visual input, hence producing more accurate and faithful responses. Extensive experimental results demonstrate that PM not only achieves superior hallucination mitigation but also enhances language generation while preserving strong reasoning capabilities. Code can be found at <https://github.com/ShunqiM/PM>.

1 Introduction

The evolution of large vision-language models (VLMs) (Liu et al., 2023; Xue et al., 2024; Wang et al., 2023; Bai et al., 2023; Yang et al., 2023b) has revolutionized the integration of computer vision and natural language processing, enabling seamless interaction between visual and textual modalities. These models effectively translate complex visual inputs into coherent and contextually relevant textual descriptions, driving advancements across a wide range of applications (Wu et al., 2023; Driess et al., 2023; Wang et al., 2022). Fueled by innovations in model architectures (Desai and Johnson, 2021; Eichenberg et al., 2022), training methodologies (Radford et al., 2021; Liu et al.,

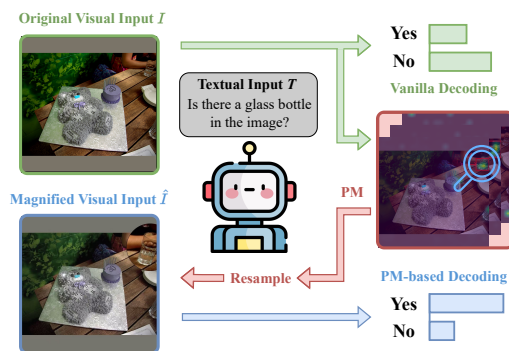


Figure 1: We propose Perception Magnifier (PM), a hallucination-mitigating decoding framework that locates and magnifies key visual regions based on inference attention patterns, enabling feature processing with finer details. For instance, PM locates the attention pattern on the top-right corner, and magnifies the glass bottle to it more discernible to the VLM.

2023; Dai et al., 2023), and diverse multimodal datasets (Schuhmann et al., 2024; Sharma et al., 2018), large VLMs continue to expand their capabilities and adaptability, solidifying their role as foundational tools in modern AI systems.

Despite their remarkable abilities, VLMs are prone to hallucination, where generated outputs fail to align with the visual input, with inaccuracies or unfaithful content. While some hallucinations arise from VLMs sharing architectural foundations with large language models (LLMs), thereby inheriting their intrinsic limitations (Liu et al., 2024b), the integration of visual and language features can also lead to hallucinations due to over-reliance on autoregressive language priors and insufficient visual representation capacity.

Various approaches have been proposed to mitigate hallucination, ranging from training-time techniques, such as debiased datasets (Liu et al., 2024a; Zhang et al., 2025b) or enlarged visual resolution (Zhai et al., 2023; Li et al., 2024), to inference-time methods such as post-processing for hallucination removal (Yin et al., 2024; Zhou et al., 2024) and specialized decoding techniques

(Li et al., 2023b; Huang et al., 2024; Leng et al., 2024). Among these, decoding-based approaches have gained particular attention for their ability to improve visual faithfulness without requiring model finetuning. These methods primarily aim to enhance the contribution of visual information during model inference to mitigate hallucination. This enhancement can be achieved explicitly: by increasing the weight of visual embeddings (Liu et al., 2025; Zhu et al., 2025), prompting key regions (Yu et al., 2024; Yang et al., 2024), or implicitly: by contrastively penalizing outputs that exhibit vague (Leng et al., 2024; Lee et al., 2024), absent (Zhang et al., 2024; Favero et al., 2024), or confused visual interpretations (Wang et al., 2024b; Chen et al., 2024a). However, these methods face still limitations in scenarios where insufficient visual resolution causes object misinterpretation.

Inspired by the observation that VLMs allocate marginal attention to many visual tokens during inference and that patch-based architectures struggle with scale-sensitive visual details, we propose the Perception Magnifier (PM), a visual decoding approach that adaptively modulates spatial resolution. As shown in Figure 1, PM magnifies critical regions for higher-resolution processing while spatially compressing rather than discarding less relevant areas, thus enhancing fine-grained detail recognition without compromising structural visual context or intrinsic reasoning ability.

Specifically, we introduce a gradient-free approach to iteratively identify visual tokens with high attention weights across layers, which are most relevant to generation, and aggregate their attention scores into a token-level heatmap. This heatmap is post-processed to enhance coverage and smoothness, then upsampled to construct a pixel-level perception map. Guided by this map, PM refines the visual input by magnifying relevant regions while compressing less critical areas, designed as a resampling process (Zheng et al., 2019; Jia et al., 2022). This dynamic adjustment allows the model to iteratively refine its focus throughout decoding, ensuring improved visual faithfulness without disrupting language reasoning.

PM is broadly applicable to interleaved VLM with unified visual sampling. Experiments on hallucination benchmarks along with GPT-4o-assisted assessments in both reference-based and vision-based settings, demonstrates that PM not only shows superior hallucination mitigation performance at question answering and open-end gen-

eration but also preserves the reasoning capabilities of VLMs. Additionally, we provide extensive ablations to analyze the contributions of each component in our method. Our contributions can be summarized as follows:

- We propose the Perception Magnifier (PM), a decoding technique that mitigates hallucination by adaptively magnifying visual regions of interest during decoding.
- We introduce a novel attention grounding method that iteratively aggregates attentions across layers to generate a perception map, which guides region magnification during decoding.
- We conduct experiments with extensive evaluations, demonstrating the effectiveness of PM in mitigating hallucination while preserving reasoning capabilities.

2 Related Work

2.1 Hallucination in Vision-Language Models

Modern vision-language models (VLMs) typically connect vision encoders with large language models (LLMs) via either cross-attention mechanisms (Li et al., 2023a; Dai et al., 2023; Alayrac et al., 2022) or projector-based token concatenation (Liu et al., 2023; Bai et al., 2023; Koh et al., 2023), with the latter dominating recent designs (Liu et al., 2024c; Hong et al., 2024; Xue et al., 2024) due to its simplicity and efficiency. While such designs align visual and textual modalities effectively, hallucination remains a persistent issue, arising not only from inherited LLM limitations (Liu et al., 2024b) but also from modality imbalance and vision encoder constraints.

To mitigate hallucination, prior efforts enhance visual grounding by prompting input images (Yu et al., 2024; Yang et al., 2024; Gizdov et al., 2024), boosting visual token influence (Zhang et al., 2024; Zhu et al., 2025; Liu et al., 2025), or improving vision module capacity (Zhai et al., 2023; Liu et al., 2024c; Chen et al., 2024c). Data-centric strategies include debiasing (Kim et al., 2025; Liu et al., 2024a; Zhang et al., 2025b; Ouali et al., 2025; Yue et al., 2024b), contrastive training (Jiang et al., 2024), preference optimization (Xie et al., 2024), evaluation techniques (Kaul et al., 2024; Guan et al., 2024; Wang et al., 2024c; Li et al., 2023c; Lovenia et al., 2024; Jing et al., 2024; He et al.,

2025), and post-hoc filtering (Yin et al., 2024; Zhou et al., 2024). While these approaches tackle hallucination from multiple angles, few methods focus on enhancing visual recognition capabilities beyond increasing resolution. We propose a decoding strategy that magnifies semantically important regions without disrupting spatial coherence, improving perception during generation.

2.2 Hallucination Mitigation at Inference

Some decoding-end approaches reduce hallucination by modifying next-token generation. Contrastive decoding (CD) (Li et al., 2023b) amplifies expert model logits while suppressing outputs from a biased source, typically a smaller or noisier model. Subsequent works further extend CD by instantiating the biased source with alternative logits, such as early-layer logits (Chuang et al., 2024) or context-unaware logits (Shi et al., 2024). OPERA (Huang et al., 2024) penalizes beam outputs that heavily attend to generic aggregation tokens. VCD (Leng et al., 2024) adapts CD for VLMs by defining biased outputs from noisy image inputs. Later works extend this by deriving biased outputs by removing image inputs (Zhang et al., 2024; Favero et al., 2024) or enhance focus on textual feature (Wan et al., 2025), masking key visual tokens (Huo et al., 2024), or synthesizing biased outputs through finetuning (Chen et al., 2024a) or prompting (Wang et al., 2024b). Others enhance expert logits by upweighting visual tokens (Liu et al., 2025; Zhu et al., 2025; Chen et al., 2025) or manipulating attentions patterns (Su et al., 2025; Prabhakaran et al., 2025). Some works explore effectively leveraging multiple outputs path (Qu et al., 2025; Lee et al., 2024). While these decoding strategies mitigate hallucination by penalizing visually inconsistent outputs, over-suppressing biased signals may hinder coherence and reasoning. Alternatively, some inference-time methods enhance visual grounding by applying masks or bounding boxes (Yu et al., 2024; Yang et al., 2024, 2023a; Chen et al., 2024b), but still struggle to recognize fine-grained details. ViCrop (Zhang et al., 2025a) improves upon this by cropping and resizing key regions as auxiliary inputs. However, it disrupts spatial structure and may impair tasks requiring holistic understanding, and doubles visual tokens’ compute. In contrast, PM magnifies semantically important regions while compressing less relevant areas, preserving spatial integrity and supporting both visual grounding and coherent generation.

3 Methods

In this paper, we propose the Perception Magnifier (PM), a novel approach that iteratively identifies and enhances key regions of the input image to refine visual input for less hallucinative token generation in VLMs. As illustrated in Figure 2, PM iteratively identifies highly attended visual tokens, blocks them, and re-forwards the sequence until no tokens receive significant attention. The identified tokens are then aggregated based on their attention scores to construct a perception map \mathcal{P} . This map is used to magnify the original image I via a sampling process, $\hat{I}^i = \mathcal{T}(I, \mathcal{P})$. The refined image \hat{I}^i then serves as the refined visual input for generating the next-step output logits, $O^{i+1} = \text{VLM}(\hat{I}^i, y_{<i})$, where $y_{<i}$ denotes the previously generated tokens. The following sections detail the perception map construction, the iterative refinement process for improving the coverage of perception map, and the attention-based magnification processed designed in the proposed PM method.

3.1 Perception Map Construction

In this section, we first introduce a simple method to construct a basic perception map to guide the magnifier. This method applies to mainstream interleaved VLMs, where patchified visual embeddings are integrated with textual tokens into a unified input sequence processed by the VLM.

The transformer processes the input sequence via self-attention. For a single attention head, the attention weight from token x_i to x_j is computed as $\text{Attn}_{i,:} = \text{Softmax}\left(Q_i K_j^\top / \sqrt{d_k}\right)$, where Q_i and K_j are the query and key vectors of tokens x_i and x_j , respectively, and d_k is the key dimension. The corresponding value V_j is then weighted by $\text{Attn}_{i,j}$ to obtain the hidden representation of x_i : $z_i = \sum_j \text{Attn}_{i,j} V_j$. In practice, transformers stack multiple layers and heads, producing attention weights across different levels. These attention patterns can be represented as a multi-dimensional tensor: $\text{Attn} \in \mathbb{R}^{N_l, N_h, H, W}$, where N_l and N_h denote the number of layers and attention heads, respectively, and $H \times W$ represents the spatial dimensions of visual tokens, which can be reshaped into a 2D feature map. PM utilizes the Attn cached during forward to construct a token-level heatmap, highlighting the most relevant visual regions based on the model attention.

Empirically, we find that attention maps from middle layers more accurately localize relevant

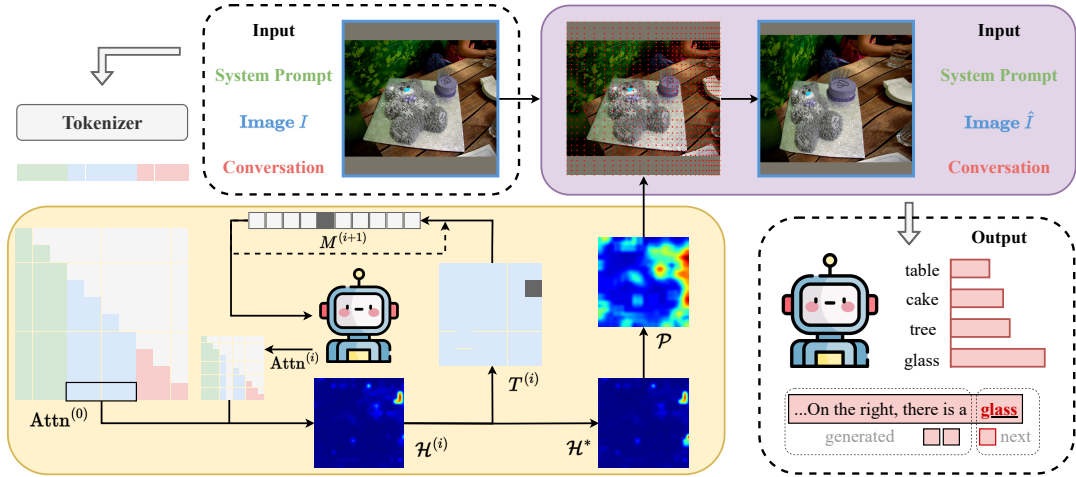


Figure 2: Structural diagram of the proposed PM method. At each decoding step, we aggregate multi-layer attentions into a token-level heatmap $\mathcal{H}^{(i)}$. By iteratively masking the most attended tokens and re-forwarding, we collect heatmaps $\{\mathcal{H}^{(i)}\}$ and combine them into \mathcal{H}^* once the process ends, and post-processing \mathcal{H}^* to obtain \mathcal{P} (yellow block). The magnifier resamples the visual input according to \mathcal{P} , producing \hat{I} (purple block). The VLM can then generate the next token with enhanced focus on critical regions with finer details.

objects than those from final layers. This aligns with recent findings showing that deep networks tend to aggregate and process visual information in early layers (Darcet et al., 2024), and that intermediate features can already support strong performance (Chuang et al., 2024). Hence, we construct the perception map by aggregating self-attention matrices from middle-to-late layers $l \geq \mathcal{L}$ on the last generated token, and applying a max pooling across all heads h to highlight the impact of more visually important regions. Formally, the token-level heatmap \mathcal{H} can be obtained by:

$$\mathcal{H} = \sum_{l=\mathcal{L}}^{N_l} \max_{h \in \{1, \dots, N_h\}} \text{Attn}_{l,h}. \quad (1)$$

Such $\mathcal{H} \in \mathbb{R}^{H,W}$ reflect the intermediate reasoning of VLM and highlight the salient regions identified.

Post-Processing to Pixel Space. The basic perception map constructed above without iterative refinement could already be upsampled to guide the magnifier. As post-processing before up-sampling the token-level heatmap \mathcal{H} to pixel-level perception map, we first refine the heatmap with methods similar to (Yu et al., 2024) to obtain a more smooth and well-distributed attention distribution for better magnification quality. Firstly, we normalize \mathcal{H} by scaling it to unit interval, obtaining \mathcal{H}' . Then, we amplify the variance of the normalized map by a scaling coefficient of α , and then compress its values back into the range $[0, 1]$ using a sigmoid function, resulting in $\mathcal{H}_{\text{proc}}$:

$$\mathcal{H}_{\text{proc}} = \text{sigmoid} \left(\alpha \cdot \frac{\mathcal{H}' - \mu(\mathcal{H}')}{\sigma(\mathcal{H}')} \right). \quad (2)$$

Finally, we apply a uniform smoothing filter with a kernel size of k to the enhanced heatmap, followed by bilinear upsampling to obtain the pixel-level perception map \mathcal{P} . As shown in Figure 2, this post-processing step refines the heatmap by amplifying low-magnitude attention values, ensuring that small but potentially important regions are also considered for magnification.

3.2 Iterative Refinement

While the basic perception map already provides a simple but effective guide for magnification, it often misses some key objects' regions when they appear multiple times. This stems from how deep vision models compress fine-grained features into a small set of tokens, commonly referred to as information registers (Darcet et al., 2024; Reddy, 2024). VLMs concentrate attention to these tokens and reduce sensitivity to semantically similar but spatially distinct regions.

To improve coverage of the perception map, we introduce an iterative refinement mechanism that progressively reveals overlooked areas by suppressing highly attended tokens from prior passes. Nonetheless, directly masking key visual regions at the image level would require fine-grained localization and re-tokenization of visual tokens, making it computationally expensive. Instead, we perform

Algorithm 1 Iterative Refinement

Require: Tokenized input (I, Q) , Visual Language Model (VLM), threshold β , max iterations T_{\max}

- 1: Initialize attention mask $M^{(0)} \leftarrow \mathbf{1}$
- 2: Initialize heatmap $\mathcal{H}^* \leftarrow \mathbf{0}$, count $\mathcal{C} \leftarrow \mathbf{0}$
- 3: **for** $i = 0$ to T_{\max} **do**
- 4: $O^{(i)}, \text{Attn}^{(i)} \leftarrow \text{VLM}(I, Q, M^{(i)})$
- 5: $\mathcal{H}^{(i)} = \sum_{l \in \mathcal{L}} \max_h \text{Attn}_{l,h}^{(i)}$
- 6: Update $\mathcal{H}^* \leftarrow \mathcal{H}^* + \mathcal{H}^{(i)}, \mathcal{C} \leftarrow \mathcal{C} + \mathbf{1}_{M^{(i)}}$
- 7: Identify $T^{(i)} \leftarrow \text{Cluster}(\mathcal{H}^{(i)})[0]$
- 8: **if** $\sum \mathcal{H}^{(i)} < \beta$ **then**
- 9: **break**
- 10: **end if**
- 11: Mask out $T^{(i)}$: $M^{(i+1)} \leftarrow M^{(i)} \cap \neg T^{(i)}$
- 12: **end for**
- 13: Normalize heatmap: $\mathcal{H}^* \leftarrow \mathcal{H}^* \div \mathcal{C}$
- 14: Post-process \mathcal{H}^* to obtain pixel-level map \mathcal{P}
- 15: **return** Refined perception map \mathcal{P}

attention blocking at the token level (Alg. 1, lines 4–5), which still blocks aggregated tokens, efficiently mitigates above challenges, and integrates easily into existing VLM pipelines.

At each iteration, we forward the input with an attention mask $M^{(i)}$, initialized as a vector of ones (line 1), and extract the token-level heatmap $\mathcal{H}^{(i)}$ (line 5). To identify visually dominant tokens $T^{(i)}$, we apply a simple 2-means clustering algorithm that partitions attention values into two groups (line 7), and select tokens in the one with the higher centroid as $T^{(i)}$. This leverages the observation that relevant tokens typically receive significantly higher attention than others. We then update the mask to exclude these tokens in the next iteration (line 11), encouraging the model to reallocate attention to previously under-attended regions.

The process continues until the total attention in the new heatmap drops below a threshold β or a maximum number of iterations is reached (lines 8–10), indicating that no additional high-attention regions remain. We aggregate attention maps across all iterations and normalize them to obtain the final refined heatmap \mathcal{H}^* (line 13), which is then post-processed to produce the pixel-level perception map \mathcal{P} (line 14). This refined map captures a broader and more accurate set of visually relevant regions, offering better guidance for magnification-enhanced decoding in different vision-language tasks.

3.3 Attention-Based Magnification

Inspired by (Zheng et al., 2019), we model the perception magnifier as an attention-based sampling process. The perception map \mathcal{P} is treated as a mass function, where regions with higher perception values are more likely to be sampled. Specifically, we first decompose \mathcal{P} into marginal-like distributions along the horizontal and vertical dimensions and approximate their cumulative distributions $\mathcal{F}_x(n)$ and $\mathcal{F}_y(n)$ as:

$$\begin{aligned}\mathcal{F}_x(n) &:= \sum_{j=1}^n \max_{1 \leq i \leq w} \mathcal{P}_{i,j}, \\ \mathcal{F}_y(n) &:= \sum_{i=1}^n \max_{1 \leq j \leq h} \mathcal{P}_{i,j},\end{aligned}\tag{3}$$

where w and h denote the width and height of the perception map, respectively. Next, we apply inverse transform sampling to remap the original image coordinates to resampled coordinates, ensuring structure-preserving magnification along both dimensions. The resulting image is constructed as:

$$\hat{I} = \mathcal{T}(I, \mathcal{P})_{i,j} = \text{Interp}(I, \mathcal{F}_x^{-1}(i), \mathcal{F}_y^{-1}(j)),\tag{4}$$

where $\mathcal{F}^{-1}(\cdot)$ denotes the inverse function of $\mathcal{F}(\cdot)$, which remaps uniformly spaced output coordinates to a perception-aware warped space. To ensure smooth spatial mapping, we use bilinear interpolation as the Interp method in the equation, balancing efficiency and effectiveness, though other interpolation methods could also be applied. The magnified image \hat{I} replaces the original visual input I and is forwarded through the VLM with the same textual input to obtain updated model output logits \hat{O} . While perception-magnified logits \hat{O} can be combined proportionally or contrastively with the original model output O for next-token generation, we directly sample from \hat{O} after the softmax function, as it already provides an enhanced representation of the visual content. The entire perception magnification process is applied at each auto-regressive decoding step, enabling the model to adaptively focus on different regions of the visual input while preserving structural integrity, hence enhancing perception ability and mitigating hallucination.

4 Experiments and Results

Datasets and Baselines. We evaluate our method on four benchmarks: MME (Fu et al., 2023), POPE (Li et al., 2023c), MMHal-Bench (Sun et al.,

	Object-level		Attribute-level		MME	MME	POPE Accuracy			
	Existence	Count	Position	Color	Percep*	Cognition	AOKVQA	GQA	COCO	AVG*
Greedy	195.00	143.33	128.33	163.33	630.00	357.86	84.10	84.39	85.29	84.59
OPERA	195.00	143.33	128.33	163.33	630.00	310.36	84.09	84.43	85.29	84.60
VCD	190.00	143.33	120.00	155.00	608.33	327.50	84.47	83.13	85.61	84.40
ICD	190.00	148.33	123.33	153.33	615.00	348.21	84.23	84.89	84.92	84.68
SID	195.00	155.00	126.67	160.00	636.67	327.86	84.39	84.17	85.81	84.79
VDD	190.00	140.00	133.33	165.00	628.33	322.64	86.73	85.52	86.71	86.32
M3ID	190.00	150.00	133.33	166.67	640.00	339.36	86.52	85.46	86.73	86.24
PAI	190.00	155.00	133.33	170.00	648.33	353.93	85.32	83.43	86.66	85.14
IBD	190.00	160.00	133.33	170.00	<u>653.33</u>	352.14	86.37	84.80	86.65	85.94
API-C	190.00	150.00	128.33	165.00	633.33	317.50	86.87	85.04	87.31	<u>86.41</u>
API-L	190.00	150.00	133.33	168.33	641.67	325.00	86.52	84.30	86.60	85.81
ViCrop-R	190.00	163.33	105.00	175.00	633.33	335.00	85.53	85.19	87.43	86.05
ViCrop-G	195.00	153.33	96.67	175.00	620.00	350.29	86.09	85.18	87.31	86.19
PM (Ours)	195.00	175.00	138.33	175.00	683.33	<u>355.36</u>	86.94	85.46	87.68	86.70

Table 1: Evaluation results for MME Perception Scores (left), Cognition Scores (mid), and POPE Accuracy (right). Columns marked with * indicate the overall hallucination scores for the benchmark. The result for BEAM is omitted for its equivalence with OPERA on single-token QA. API and ViCrop each propose multiple variants, denoted by postfixes (e.g., -C or -L), with details provided in Section B.

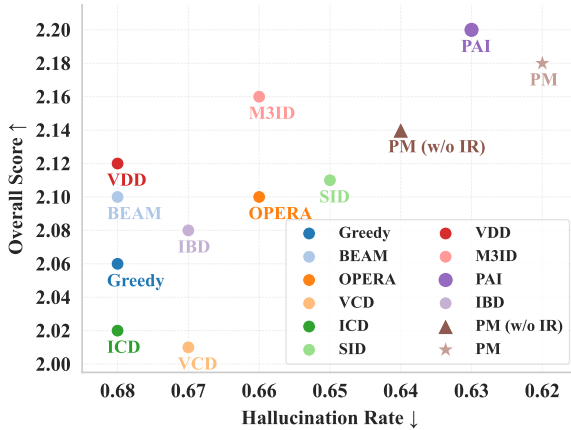


Figure 3: MMHal-Bench comparison of hallucination mitigation decoding methods. Each point represents one method, with lower hallucination rate and higher overall score indicating better performance.

2024) and LLaVA-Bench (Liu et al., 2023), all commonly used for assessing multimodal hallucination. For comparison, we include various decoding or inference strategies, including greedy decoding, VCD (Leng et al., 2024), VDD (Zhang et al., 2024), M3ID (Favero et al., 2024), IBD (Zhu et al., 2025), PAI (Liu et al., 2025), ICD (Wang et al., 2024b), SID (Huo et al., 2024), beam search (Sutskever et al., 2014), OPERA (Huang et al., 2024), API (Yu et al., 2024), and ViCrop (Zhang et al., 2025a). Additional methods such as (Yue et al., 2024b; Shi et al., 2024), which rely on contextual information or EOS decisions, are less relevant to visual hallucination and are therefore excluded. Further details are provided in Sections A and B.

Implementation Details. We conduct our experiments using LLaVA-1.5 7B (Liu et al., 2024c). We start aggregating attentions from layer $\mathcal{L} = 12$, with scaling coefficient and kernel size $\alpha = 10$, $k = 3$, and attention threshold for iterative refinement $\beta = 0.3$.

4.1 Quantitative Analysis

Table 1 reports results on MME and POPE. POPE centers on existence-based queries, where most methods benefit from reinforcing visual token importance. MME is more challenging, with its perception subset targeting hallucination and its cognition subset requiring higher-level reasoning. While our method focuses on mitigating hallucination, we report cognition results for completeness, noting that most baselines underperform on this subset as modifications on the embedding and logit could harm the VLM reasoning.

Among baselines, CD-based methods suppress hallucinated logits but often fail when visual detail remains ambiguous across both logits. Embedding-level methods such as PAI and IBD improve perception by reinforcing visual features across multiple layers. However, they may still struggle when relevant regions are too small or diffuse to be effectively represented in the ViT backbone. API suppresses irrelevant regions via masking, but does not enhance salient visual content. ViCrop isolates key areas via cropping, which may hinder spatial understanding due to loss of context and potential ambiguity from dual-image inputs, resulting in limited MME gains. In contrast, our method

LLaVA-Bench	COCO			In-the-Wild		
	Ref \uparrow	Acc \uparrow	Det \uparrow	Ref \uparrow	Acc \uparrow	Det \uparrow
Greedy	81.15	92.69	95.76	56.94	71.83	79.95
BEAM	83.17	94.75	91.60	58.94	74.71	78.25
OPERA	82.99	96.12	93.61	58.44	71.81	77.92
VCD	82.00	92.21	94.39	57.17	69.4	81.22
ICD	83.29	93.91	95.46	54.43	66.89	74.81
SID	81.10	<u>97.35</u>	95.80	59.49	70.93	81.63
VDD	81.38	95.62	102.75	61.60	78.22	83.01
M3ID	83.02	96.41	95.78	<u>61.88</u>	77.49	82.83
PAI	79.04	92.06	90.59	52.29	69.74	72.49
IBD	<u>85.29</u>	95.73	95.45	61.59	<u>77.91</u>	82.86
PM (Ours)	85.76	97.74	<u>95.80</u>	62.46	<u>77.21</u>	84.50

Table 2: Results on LLaVA-Bench (%). Ref denotes reference agreement, while Acc and Det represent accuracy and detailedness in vision-based evaluation.

magnifies the relevant regions while preserving surrounding context and internal states of the VLM. This preserves the embedding distribution and enhances fine-grained recognition without compromising spatial coherence or reasoning, leading to state-of-the-art results across both benchmarks.

We further evaluate on MMHal-Bench in Figure 3, which provides a complementary setting for hallucination assessment in open-end generation. On this benchmark, PM is competitive with the strongest baseline in overall score while achieving the lowest hallucination rate among all compared methods. We attribute the relatively modest gain in overall score to the design of MMHal-Bench, where many questions concern attributes of the primary object in the scene. In such cases, the benefit of magnifying small or cluttered regions is naturally less pronounced. Nevertheless, PM still consistently reduces hallucinations, and the full model further improves over PM w/o IR on both overall score and hallucination rate, supporting the effectiveness of iterative refinement.

In addition, all methods exhibit varying degrees of degradation on the MME cognition subset, whereas ours does not. A detailed analysis of how and why different methods affect the cognitive ability is provided in Section E.

4.2 GPT-4o Assisted Evaluation

Previous studies (Huang et al., 2024; Zhu et al., 2025) have demonstrated the reliability of GPT-4-assisted evaluation for hallucination assessment. We extend their evaluation protocol and experiment on LLaVA-Bench to benchmark the open-end generation quality. Overall, GPT-4o scores responses based on reference agreement (*Ref*), visual accuracy (*Acc*), and descriptive detail (*Det*), depending on the evaluation setting. We omit API and ViCrop

MME-Perception	Object-level		Attribute-level		Percep Total*
	Existence	count	position	color	
Greedy	195.00	143.33	128.33	163.33	630.00
PM w/o IR & MLA	195.00	155.00	120.00	170.00	640.00
PM w/o MLA	195.00	155.00	125.00	170.00	645.00
PM w/o IR	190.00	175.00	135.00	165.00	665.00
PM	195.00	175.00	138.33	175.00	683.33

Table 3: Ablations on iterative refinement (IR) and multi-layer aggregation (MLA). Both contribute to improvements compared to their respective baselines.

as they are not applicable to open-ended generation. Details on the GPT-4o metrics and prompts can be found in Section K.

As shown in Table 2, PM achieves the most competitive overall performance on open-end generation tasks. While its detailedness on COCO and accuracy on the In-the-Wild split are slightly lower than a few methods, PM still demonstrates advantage in hallucination mitigation, particularly in reference-based evaluation. On COCO, the detailedness gap appears to be mainly related to response verbosity rather than clearly inferior visual grounding. On the In-the-Wild split, the relatively weaker gains are mostly observed in cases where answers rely more on language priors or world knowledge than on fine-grained visual details, which are less aligned with the strength of PM. These results indicate that even in open-end decoding, where the model does not always need to focus on specific regions at each step, the proposed adaptive perception magnifier can still enhance output quality and effectively reduces hallucinations.

4.3 Qualitative Analysis

Figure 4 shows qualitative examples depicting the input question, original image, raw and processed perception map \mathcal{H}^* and \mathcal{P} , and the perception-magnified image. The results demonstrate that the proposed method effectively enlarges small objects relevant to the question while preserving coarse-grained structural information. In the first example, the VLM initially attends to the top-left region, recognizing the potentially relevant region but still failing to identify the chair. This suggests that accurate attention does not necessarily imply correct recognition, as local evidence may still be insufficient under limited resolution. With the magnified input, the chair is presented at a higher spatial resolution, enabling the VLM to correctly recognize it. This illustrates that accurate attention does not imply recognition: local evidence can be overridden under limited resolution. With the magnified input,

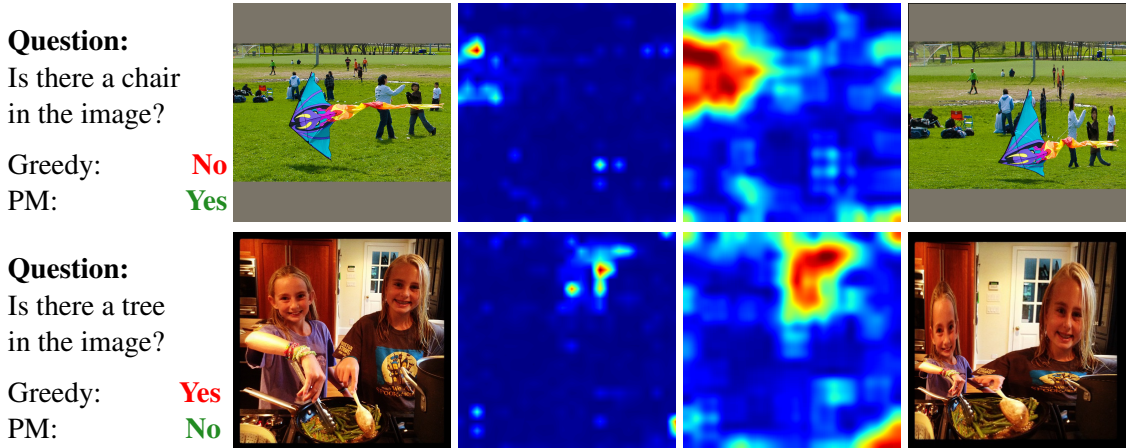


Figure 4: Qualitative examples showing the input question, the original image, the raw and processed perception map \mathcal{H}^* and \mathcal{P} , and the perception-magnified image, arranged from left to right. The answers generated by greedy decoding and PM are compared, with correct answers highlighted in green and incorrect ones in red.

the chair is presented at a higher spatial resolution, enabling the VLM to correctly recognize it. Similarly, in another case, the VLM misidentifies an object behind a door window as a tree. However, the perception-magnified input allows the model to re-examine the object with greater detail, preventing hallucinated predictions.

Overall, the qualitative results suggest that the proposed method enhances the perceptibility of fine-grained visual details in regions critical for next-token generation. By enabling the model to examine these areas at a higher resolution, our approach effectively reduces hallucinations and improves overall model performance.

4.4 Ablations

Iterative Refinement. We ablate the effect of iterative refinement in constructing the perception map. As shown in Table 3, the refined maps consistently outperform their non-iterative counterparts, confirming the benefit of this strategy in improving visual localization. Iterative refinement helps address tokens overlooked from the early-stage aggregation of visual features, resulting in more accurate coverage of relevant regions. Although some residual leakage from masked tokens remains due to information exchange in the ViT (Dosovitskiy et al., 2021) visual feature extractor, our approach still achieves decent grounding of key areas with minimal computational overhead, making it an effective and efficient solution for guiding the magnifier.

Multi-Layer Aggregation. We ablate the effect of aggregating attention across multiple layers when constructing the perception map. As shown

MME-Perception	Object-level Existence	Count	Attribute-level Position	Color	Percep Total*
<i>(a) Perception Map Construction Methods</i>					
API CLIP	190.00	155.00	131.67	170.00	646.67
API LLaVA	190.00	165.00	126.67	170.00	651.67
ViCrop-R	195.00	163.33	116.67	170.00	645.00
ViCrop-G	195.00	158.33	116.67	170.00	640.00
Rel Map	195.00	175.00	130.00	155.00	655.00
PM w/o IR	190.00	175.00	135.00	165.00	665.00
<i>(b) Perception-Guided Magnification Methods</i>					
Blurring	185.00	158.33	111.67	175.00	630.00
Bounding Box	180.00	160.00	130.00	170.00	640.00
Masking	195.00	155.00	128.33	170.00	648.33
ViCrop	195.00	153.33	123.33	175.00	646.67
DualView	190.00	155.00	135.00	170.00	650.00
Magnification	195.00	175.00	138.33	175.00	683.33

Table 4: Ablation results across two groups: (a) comparing different methods for constructing perception maps; (b) comparing prompting methods that highlight key regions based on the map. IR denotes iterative refinement.

in Table 3, multi-layer aggregation yields significantly better performance compared to using attention from a single layer. While middle-to-deep layers generally focus more on visual content, the most informative attention varies across layers and samples. In many cases, only a subset of intermediate layers exhibits structured visual attention, while others focus primarily on sink tokens (Xiao et al., 2024). Aggregating across layers reduces this variability, ensuring that key visual cues are reliably captured for guiding the magnifier.

Comparison with Different Maps. Attention-based visualization methods (Chefer et al., 2021; Stan et al., 2024) have been proposed to improve VLM interpretability. We adapt several representative techniques, including CLIP-based Rel(evance) Maps, API (Yu et al., 2024), and ViCrop (Zhang

et al., 2025a), to construct a map to guide the magnifier, and compare them to our approach w/o IR using identical post-processing. As shown in Table 4, our method achieves superior coverage of relevant visual regions. Attribution maps based on CLIP often reflect global similarity rather than token-level saliency, while API and ViCrop variants aggregate attention from a single layer, resulting in lower stability and diluted focus. Backpropagation-based methods (e.g., Rel Map, ViCrop-G) offer fine-grained attribution but contribute little to magnification quality and are computationally expensive. In contrast, the proposed strategy, even w/o the IR process, preserves critical visual cues with better spatial consistency and minimal overhead, making it well-suited for efficient perception-guided decoding.

Image Prompting Methods. Various methods have been proposed to “prompt” images to emphasize specific regions (Yang et al., 2024; Zhang et al., 2025a), including adding a bounding box, applying blurring effect, masking out the remaining regions, and cropping out the key area (ViCrop), or append the magnified image, as dual inputs (DualView). We benchmark these methods as alternatives to our magnifier, evaluating their effectiveness in highlighting key visual regions when combined with the constructed perception map. As shown in Table 4, our magnification approach more effectively leverages the perception map to enhance the perception ability of the VLM.

Concretely, magnification not only highlights important regions but also reveals fine-grained details within limited resolution, allowing the model to capture subtle visual patterns. In contrast, many prompting methods merely emphasize region importance without increasing effective resolution, limiting their impact on fine-grained recognition. Meanwhile, ViCrop also yields certain improvement, as it also enhances local details with resized key regions. However, ViCrop discards contextual information and confuses the VLM by supplying the additional resized image, leading to reduced effectiveness in tasks such as position-related and counting questions. The DualView methods

5 Conclusion

We introduce the Perception Magnifier (PM), a visual decoding method that refines image inputs to mitigate hallucinations in vision-language models. PM adaptively enlarges visually critical regions

identified by multi-layer attention iteratively, providing enhanced effective resolution on relevant finer-grained details at each decoding step. Experiments demonstrate that PM significantly reduces hallucinations at VQA and open-end generation, while preserving reasoning capabilities. Future work will focus on optimizing the efficiency for PM, and mitigate shape distortions.

Limitations

Although PM effectively mitigates hallucinations, it also introduces several limitations. Region magnification may distort local shapes, reducing reliability in tasks requiring geometric fidelity. The method further incurs computational overhead, as magnification interrupts efficient decoding strategies like KV caching. In addition, applying PM to models with complex token-image mappings can require additional attention alignment mechanisms. Further detailed discussions are available at Section J.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 23716–23736.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.
- Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.
- Beitao Chen, Xinyu Lyu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. 2024a. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*.
- Xinlong Chen, Yuanxing Zhang, Qiang Liu, Junfei Wu, Fuzheng Zhang, and Tieniu Tan. 2025. Mixture of decoding: An attention-inspired adaptive decoding

- strategy to mitigate hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8525–8542.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024b. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 7824–7846.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 49250–49267.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision transformers need registers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11162–11173.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: an embodied multimodal language model. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8469–8488.
- Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. MAGMA – multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP (Findings: EMNLP)*, pages 2416–2428.
- Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Andrey Gizdov, Shimon Ullman, and Daniel Harari. 2024. Seeing more with less: Human-like representations in vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoub, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. 2025. Evaluating and mitigating object hallucination in large vision-language models: Can they still see removed objects? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6841–6858.
- Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, and 6 others. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.

- Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. 2022. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4722.
- Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaying Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27036–27046.
- Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP (Findings: EMNLP)*, pages 5042–5063.
- Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27228–27238.
- Minchan Kim, Minyeong Kim, Junik Bae, Suhwan Choi, Sungkyung Kim, and Buru Chang. 2025. Exploiting semantic reconstruction to mitigate hallucinations in vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 236–252.
- Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Yi-Lun Lee, Yi-Hsuan Tsai, and Wei-Chen Chiu. 2024. Delve into visual contrastive decoding for hallucination mitigation of large vision-language models. *arXiv preprint arXiv:2412.06775*.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12286–12312.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26763–26773.
- Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 34892–34916.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2025. Paying more attention to image: A training-free method for alleviating hallucination in vlms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 125–140.
- Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2024. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. In *Proceedings of the Workshop on Advances in Language and Vision Research (ALVR)*, pages 37–58.
- Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2025. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in vlms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 395–413.

- Vishnu Prabhakaran, Purav Aggarwal, Vinay Kumar Verma, Gokul Swamy, and Anoop Saladi. 2025. VADE: Visual attention guided hallucination detection and elimination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14949–14965.
- Xiaoye Qu, Jiashuo Sun, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. 2025. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. In *International Conference on Computational Linguistics (COLING)*, pages 4428–4441.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763.
- Gautam Reddy. 2024. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4035–4045.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2024. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernamed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.
- Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwala, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: Explainable AI for Computer Vision Workshop (CVPRW)*.
- Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 3104–3112.
- Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. 2025. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:95095–95169.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 23318–23340.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 121475–121499.
- Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL (Findings: ACL)*, pages 15840–15853.
- Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. 2024c.

- Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 288–304.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273.
- Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgankar, Shelby Heinecke, and 8 others. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023a. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.
- Lingfeng Yang, Yuezhe Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2024. Fine-grained visual prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 24993–25006.
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *arXiv preprint arXiv:2309.17421*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.
- Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024. Api: Attention prompting on image for large vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, page 251–268.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.
- Zihao Yue, Liang Zhang, and Qin Jin. 2024b. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11766–11781.
- Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. 2023. Halle-switch: Controlling object hallucination in large vision language models. *arXiv preprint arXiv:2310.01779*.
- Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025a. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2025b. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 196–213.
- Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. Debiasing large visual language models. *arXiv preprint arXiv:2403.05262*.
- Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 2019. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5012–5021.
- Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025. IBD: Alleviating hallucinations in large vision-language models via image-biased decoding. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Through the Magnifying Glass: Adaptive Perception Magnification for Hallucination-Free VLM Decoding

Appendix

A Datasets

We carry out our experiments on three benchmarks popular for examining the perception ability and hallucination of VLMs in English.

MME (Fu et al., 2023) benchmark is designed to comprehensively assess both perception and cognition abilities of VLMs across 14 distinct subtasks on 2,370 questions. In our study, we focus primarily on the *Perception* subset of MME following (Leng et al., 2024) to evaluate the effectiveness of our method in controlling hallucinations.

POPE (Polling-based Object Probing Evaluation) (Li et al., 2023c) is a $9 \times 3,000$ question benchmark for evaluating object hallucination in VLMs. It formulates hallucination evaluation as a binary classification task by posing yes/no questions about object presence, addressing limitations of CHAIR (Rohrbach et al., 2018), which relies on exact text matching, and providing a more robust evaluation less sensitive to prompt variations.

LlaVA-Bench (Liu et al., 2023) is a benchmark designed to evaluate VLMs in real-world visual chat scenarios. It comprises two components: LLaVA-Bench (COCO), utilizing images from the COCO dataset, and LLaVA-Bench (In-the-Wild), a more challenging subset featuring 24 diverse images, such as indoor and outdoor scenes, memes, paintings, and sketches, paired with 60 detailed, manually curated questions. The benchmark covers various vision-language tasks, making it suitable for assessing model generalization. In our study, we employ both COCO and In-the-Wild splits to comprehensively evaluate different methods on hallucination control, cognitive abilities, and performance in open-ended question answering.

We also evaluate our method and baselines on **MMHal-Bench** (Sun et al., 2024), a multimodal hallucination benchmark consisting of 96 open-ended image-question pairs, spanning diverse 8 question types and 12 object categories. MMHal-Bench adopts a GPT-4-based evaluation protocol that measures both average score and hallucination rate for VLM-generated outputs. We do not include additional benchmarks such as MMMU (Yue et al., 2024a) and MathVision (Wang et al., 2024a), as

they are dominated by chart-based and symbolic problems where failures are primarily due to abstract reasoning rather than hallucination or misperception, and are thus largely orthogonal to our focus.

B Baselines

This section summarizes the high-level idea of methods used as comparisons in the experiments. We apply these methods on our benchmarks with the default hyperparameters provided in the original paper.

Beam Search explores multiple output candidates by iteratively expanding the most probable sequences and selecting the top- k hypotheses at each step, resulting in improved answer diversity but at the cost of increased computational complexity proportional to the beam size. We evaluated all beam-based methods with a beam size of 4.

OPERA (Huang et al., 2024) is a beam-search-based method, examining the presence of a hallucinate beam and pruning the beam based on its attention on information aggregation patterns.

VCD (Leng et al., 2024) introduces noise into the input image and processes it through the VLM to generate a hallucinative output. By leveraging contrastive decoding (CD) to penalize high-probability predictions that align with the hallucinative output, VCD reduces the likelihood of hallucinations in the final prediction, resulting in more reliable responses.

ICD (Wang et al., 2024b) generates a hallucinative output for CD-based decoding by appending “*You are a confused object detector*” to the input prompt.

SID (Huo et al., 2024) induces a hallucinative output for CD-based decoding by identifying key visual tokens in the image and pruning them through forward.

VDD (Zhang et al., 2024) mitigates language-induced hallucinations in VLMs by generating a hallucinative output through the removal or replacement of visual tokens with <unk> tokens.

M3ID (Favero et al., 2024) is similar to VDD but with weighted strength of incorporating contrastive corrections based on the text-image mutual infor-

Method	AOKVQA			GQA			COCO		
	Adversarial	Popular	Random	Adversarial	Popular	Random	Adversarial	Popular	Random
Greedy	78.67	84.93	88.70	80.40	83.27	89.50	83.50	85.57	86.80
Opera	78.60	84.93	88.73	80.43	83.33	89.53	83.47	85.57	86.83
VCD	79.10	84.93	89.37	79.63	81.70	88.07	83.53	85.63	87.67
ICD	79.07	85.40	88.23	80.97	84.77	88.93	83.20	85.23	86.33
SID	78.90	85.40	88.87	80.60	83.33	88.57	83.90	86.07	87.47
VDD	81.03	88.27	90.90	82.00	85.27	89.30	85.00	87.00	88.13
M3ID	80.87	87.97	90.73	81.67	84.93	89.77	85.10	86.93	88.17
PAI	79.63	86.27	90.07	79.90	82.30	88.10	85.03	86.97	87.97
IBD	80.60	87.47	91.03	81.30	84.10	89.00	84.90	86.97	88.07
API-C	81.00	88.20	91.40	81.33	84.27	89.53	85.47	87.53	88.93
API-L	80.80	87.70	91.07	81.00	83.57	88.33	84.53	86.97	88.30
ViCrop-R	80.00	86.37	90.23	81.60	83.83	90.13	85.60	87.70	89.00
ViCrop-G	80.97	86.77	90.53	81.67	83.93	89.93	85.13	87.70	89.10
PM (Ours)	81.07	87.77	92.00	81.73	84.37	90.30	85.87	88.00	89.17

Table 5: Detailed evaluation results for POPE accuracy with LLaVA 1.5.

mation.

PAI (Liu et al., 2025) enhances visual faithfulness by increasing the weights of visual tokens through attention mechanisms while penalizing outputs generated without visual inputs using CD.

IBD (Zhu et al., 2025) also biases the output toward visual inputs by adding visual attentions and refines it against the original output using CD. Additionally, it adaptively adjusts the contrastive update strength at each decoding step based on contextual needs.

API (Yu et al., 2024) highlights key visual regions by overlaying a gray-level mask on the input image. The mask is guided by a heatmap derived from image-text associations using a CLIP model (-C) or a LLaVA model (-L).

ViCrop (Zhang et al., 2025a) estimates a bounding box around the key region in the image and generates a new input by cropping and resizing this region. Three approaches are proposed for bounding box estimation with different level of reliance on gradient information: relative attention (-R), gradient attention (-G), and pure gradient (-P). The pure-gradient approach is not benchmarked due to its higher GPU memory consumption, which necessitates lower model precision and results in degraded performance, making it unsuitable for a fair comparison.

C Computing Infrastructure

All experiments were conducted on a single machine equipped with an NVIDIA GeForce RTX 3090 GPU and an Intel(R) Core(TM) i7-10700 CPU. The system runs Ubuntu 22.04 LTS and uses CUDA version 12.1 for GPU acceleration. Experiments were implemented in Python using PyTorch

MME-Cognition	CS	NR	Trans	CR	Total*
Greedy	112.86	70.00	107.50	67.50	357.86
OPERA	112.86	47.50	90.00	60.00	310.36
VCD	120.00	37.50	97.50	72.50	327.50
ICD	120.71	70.00	92.50	65.00	348.21
SID	122.86	55.00	85.00	65.00	327.86
VDD	110.14	58.00	68.00	86.00	322.64
M3ID	110.86	61.00	92.50	75.00	339.36
PAI	116.43	70.00	100.00	67.50	353.93
IBD	117.14	70.00	107.50	57.50	352.14
API-C	115.00	47.50	107.50	47.50	317.50
API-L	120.00	45.00	115.00	45.00	325.00
ViCrop-R	115.00	50.00	105.00	65.00	335.00
ViCrop-G	109.29	50.00	105.00	86.00	350.29
PM (Ours)	117.86	42.50	107.50	87.50	355.36

Table 6: Comparison of different hallucination control decoding methods on the MME-Cognition subset, including Commonsense Reasoning (CS), Numerical Reasoning (NR), Translation (TR), and Code Reasoning (CR).

(v2.4.0) and Transformers (v4.37.2). Experiments on all benchmarks take around 24 GPU hours in total.

D License of Artifacts

We use the following resources in accordance with their respective licenses and terms: MME (no explicit license declared at the time of writing), POPE (MIT License), LLaVA-Bench (Apache 2.0 License), the LLaVA model (Apache 2.0 License), and Shikra (CC-BY-NC 4.0 License). All resources are employed strictly for research purposes, and we do not redistribute them beyond what is permitted by their original licenses and terms of use.

E Maintaining Cognition Ability

Many hallucination control decoding methods enhance perception by increasing the contribution of visual information during token generation. While

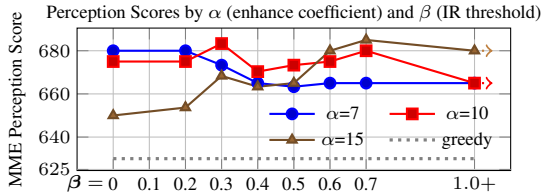


Figure 5: Hyperparameter analysis of α and β on MME. PM demonstrates robust performance across different settings.

these approaches generally improve perceptual ability, they often disrupt the calibrated balance between the vision and language modalities, potentially compromising the inherent cognitive capabilities learned by the VLM during large-scale pre-training.

To investigate this, we evaluate our method and baselines on the Cognition split of MME, where answering requires reasoning beyond direct visual cues. As shown in Table 6, most CD-based methods that explicitly modify output logits exhibit a notable decline in cognition ability, aligning with expectations. Surprisingly, PAI and IBD, which enhance the attention weight of visual tokens through intermediate embedding modifications, do not incur a substantial loss in cognition performance. This suggests that the bias introduced by their layer-wise manipulation is relatively mild and can be compensated for by subsequent model layers. Furthermore, despite API-based methods modify the image without altering the VLM’s internal reasoning, they still lead to degraded performance. We hypothesize that the gray masking applied to images in the cognition set where backgrounds are typically uniform may introduce additional perceptual burden, ultimately impacting overall performance.

In contrast, our proposed PM method does not negatively affect cognition. Although our approach is designed solely for hallucination control rather than cognition enhancement, the total cognition score remains comparable to greedy decoding and outperforms other methods, demonstrating its robustness in preserving reasoning capabilities.

F Hyperparameter Analysis on α and β

We analyze the effect of α and β , which control magnification strength and the stopping criterion for iterative refinement, respectively. Increasing α enhances heatmap contrast and strengthens magnification, while decreasing β increases refinement steps, improving coverage but at higher computa-

tional cost (capped at 8 iterations). As shown in Figure 5, smaller α pairs well with low β , which already ensures broad region coverage, reducing the need for strong magnification. Overall, the perception magnifier remains robust across a wide range of hyperparameters, consistently identifying key visual regions. Notably, using a larger β approximates the performance of the basic PM w/o iterative refinement, yet substantially mitigates hallucinations with only $2\times$ the compute.

G Quantitative Results on POPE

Table 5 reports detailed POPE results across 9 splits. Overall, PM achieves the best performance on most splits and attains the strongest overall average, indicating that perception-guided magnification is effective for mitigating object hallucination in yes/no existence-based VQA. Notably, the absolute gain in AVG is modest for all training-free methods, because POPE accuracies with LLaVA-1.5 are already relatively high and thus largely bounded by the fixed backbone.

We observe that the GQA-popular splits yield smaller improvements for PM. This subset tends to contain denser scenes with lower object saliency, making attention-guided magnification less decisive. In addition, popular questions are more susceptible to language-prior bias toward frequent object categories, where contrastive debiasing methods (e.g., VDD/M3ID) can be slightly stronger by explicitly suppressing such priors. In contrast, PM primarily enhances perceptual evidence via magnification while keeping the model’s internal vision-language calibration intact. Despite this, PM remains competitive on the popular splits and consistently improves on the remaining splits, demonstrating robust generalization across datasets and question distributions.

H Performance on Shikra

To assess the generalizability of PM beyond LLaVA, we apply it, along with several comparative methods, to Shikra-7B (Chen et al., 2023), a VLM designed for referential dialogue. The performance is evaluated on MME and POPE using Shikra with iterative refinement threshold $\beta = 1.5$. The results are presented in Table 7. Note that certain comparative methods are excluded either due to the lack of available implementations for Shikra or constraints on the computational resources available.

	Object-level		Attribute-level		MME Score Total*	POPE Accuracy			
	Existence	Count	Position	Color		AOKVQA	GQA	COCO	AVG*
Greedy	190.00	160.00	71.67	155.00	576.67	81.87	81.27	83.46	82.19
VCD	185.00	163.33	65.00	148.33	561.67	81.86	81.14	83.31	82.10
ICD	195.00	165.00	65.00	145.00	575.00	81.93	81.05	83.78	82.25
SID	195.00	155.00	73.33	145.00	568.33	81.57	81.40	83.43	82.13
VDD	195.00	155.00	73.33	155.00	578.33	81.99	81.74	83.03	82.26
M3ID	195.00	160.00	75.00	155.00	585.00	81.92	81.54	83.43	82.30
API-C	185.00	165.00	76.67	165.00	591.67	82.21	81.28	83.47	82.32
API-L	190.00	165.00	73.33	160.00	588.33	81.84	81.24	83.33	82.14
PM (Ours)	190.00	165.00	68.33	190.00	613.33	82.15	81.27	83.66	82.36

Table 7: Evaluation results for MME Perception Scores (left) and POPE Accuracy (right) on the **Shikra** VLM. Columns marked with * indicate the overall scores for the benchmark. Some methods are omitted due to limitations on computational resources and implementations. The proposed PM method consistently demonstrates superior performance on both benchmarks.

As shown in Table 7, the proposed method continues to improve performance on Shikra compared to the baselines, though the improvement is less pronounced than on LLaVA. This discrepancy may stem from Shikra’s lower visual resolution and fewer visual tokens, which impact the effectiveness of perception magnification. With fewer visual tokens, cases where an object is partially recognized at intermediate layers but not reflected in the final prediction become less frequent, as each token carries greater attention weight in the inference process. Therefore, the proposed method has fewer opportunities to enhance recognition. Additionally, the reduced number of visual tokens lowers the granularity of the token-level perception map, leading to less precise perception magnification, which may further limit the performance gains of PM.

This observation suggests that the proposed method could be particularly beneficial for high-resolution VLMs, where small local features are more likely to be detected at intermediate layers but fail to significantly influence the final prediction. However, verifying this hypothesis is constrained by computational resource limitations.

Overall, the results on Shikra provide strong evidence that the proposed PM method generalizes well beyond LLaVA and effectively enhances hallucination mitigation in applicable VLMs.

I Qualitative Samples of the PM

Figure 7 presents additional qualitative examples illustrating the input question, the original image, the processed perception map \mathcal{P} , and the perception-magnified image. The results demonstrate that the proposed method effectively enlarges small objects relevant to the question while preserving fine details. In many cases, the VLM benefits from the

perception-magnified image, as it can now recognize previously overlooked objects that were too small to be detected in the original image. This leads to more accurate predictions, as observed in the first and second rows.

Furthermore, when the VLM erroneously identifies a non-existent object as present, as shown in the third example, PM magnifies the corresponding region, enabling the model to reassess the scene with finer detail. This process helps reduce false positive hallucinations by magnifying the visual details hence improve perceptibility. However, localized magnification does not always guarantee improved visual faithfulness. As demonstrated in the fourth row, despite providing a closer view, the model now fails to recognize the cabinet in the background. This may be due to the cabinet being identified based on its structural context rather than fine-grained textures, highlighting the limitations of region-focused magnification.

Overall, these results suggest that the proposed method effectively enhances the perceptibility of visual detail on regions crucial for next-token generation while generally preserving the overall structural integrity of objects, ultimately reducing hallucinations and improving model performance.

J Limitations and Future Works

While the proposed PM method improves VLM inference by mitigating hallucinations, magnifying image regions may also introduce certain drawbacks. This section discusses key limitations of our approach.

Shape Distortion in Region Magnification. Although region magnification allows the model to observe fine-grained visual details without alter-

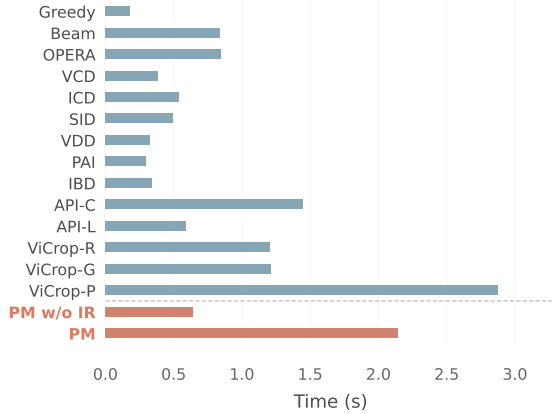


Figure 6: Average per-question running time (s) for different decoding methods.

ing the overall image structure, it may distort local shape properties. Specifically, in tasks requiring precise shape or length recognition, such as distinguishing letters or assessing object sizes, PM does not guarantee proportional magnification. This can introduce challenges for accurate recognition and induces hallucinations. For instance, when applied to CLEVR-style questions, as shown in Figure 8, with the query “Please estimate how many times large is the blue sphere compared to the red cube?”, PM magnifies both objects, causing the blue sphere to appear visually larger while distorting the shape of the other two cubes. Such unintended modifications may negatively impact final predictions. Future work could explore refinements in perception map construction or post-processing techniques to achieve more structure-preserving magnification. To address the issue of distortion, several enhancements can be integrated into our framework. One potential direction involves incorporating detected bounding boxes to constrain the perception map, thereby regularizing the magnification process and preserving object shape integrity. Another approach is to group visual tokens with shared semantic representations, enabling uniform resampling across coherent regions and reducing uneven deformation within individual objects. Furthermore, for tasks where geometric fidelity is critical such as those involving precise reasoning over shape or size, magnification can be applied adaptively at selective decoding steps rather than continuously, allowing better control over the trade-off between visual detail enhancement and structural preservation.

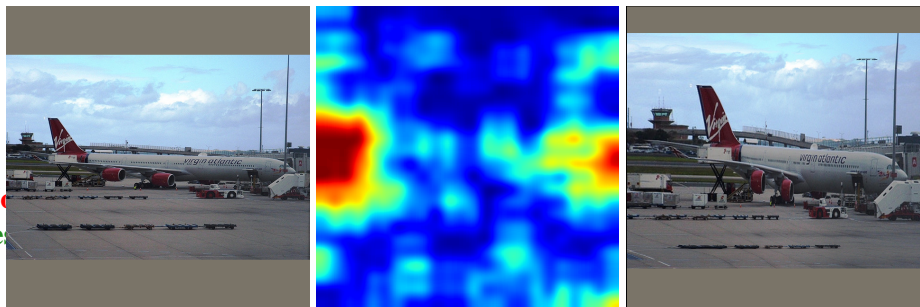
Efficiency. A key limitation of our method lies in its computational overhead: applying perception-based magnification at every decoding step requires feature extraction from a modified image, which precludes the use of efficient VLM decoding techniques such as key-value (KV) caching. As depicted in Figure 6, PM without Iterative Refinement (IR) is approximately $3.5\times$ slower than greedy decoding, while PM with IR increases this to $\sim 12\times$. In comparison, contrastive-decoding baselines such as IBD and VDD incur only $\sim 2\text{--}3\times$ inference time, whereas prompting-based methods with multi-pass processing (e.g., API and ViCrop) typically cost $8\text{--}20\times$ depending on the variant. These results suggest that PM exposes a controllable cost–performance trade-off: PM w/o IR already yields substantial gains at moderate overhead, while full PM can be reserved for latency-tolerant settings.

Nonetheless, our method remains relatively efficient among perception-oriented approaches when used without iterative refinement, as the perception map can be constructed with only a single additional VLM forward pass. To further mitigate runtime cost, PM can be applied adaptively only at decoding steps where visual information is critical (e.g., during object description), while being omitted for tokens with lower visual dependency such as punctuation or purely reasoning-oriented connectives. A practical indicator for such adaptation is the aggregated attention mass on visual tokens, which signals when the model is actively grounding its predictions in the image. Finally, future work may explore speculative or deferred refinement strategies, where magnification is applied post hoc to selectively revise previously decoded spans, thereby reducing the frequency of image resampling and improving scalability for longer generations.

Improving Compatibility. While the proposed method can be easily applied to interleaved VLMs with visual tokens corresponding to all patch features, applying it to VLMs where the correlation between visual tokens and image regions are non-trivial, can be slightly more complicated. The method depends on certain mapping from the visual token attention weights to the attention span on the original image. Therefore, when apply PM to models where the visual tokens are constructed with not only a feature extractor, but also a token sampler or cross-attention feature extractors, PM

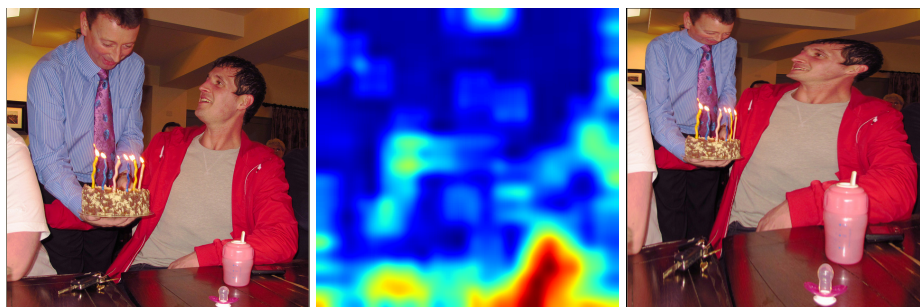
Question:
Is there a building
in the image?

Greedy: **No**
PM: **Yes**



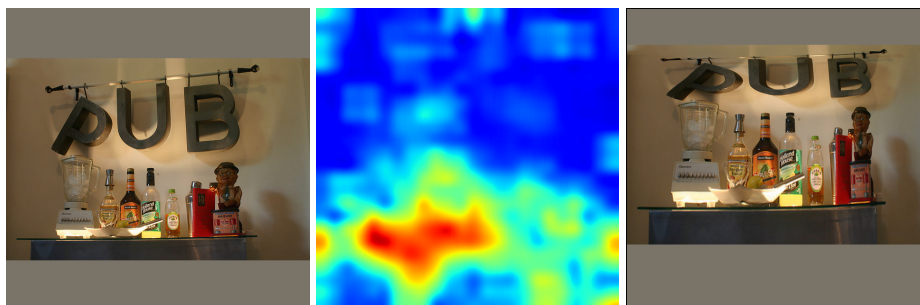
Question:
Is there a bottle
in the image?

Greedy: **No**
PM: **Yes**



Question:
Is there a paper
in the image?

Greedy: **Yes**
PM: **No**



Question:
Is there a cabinet
in the image?

Greedy: **Yes**
PM: **No**

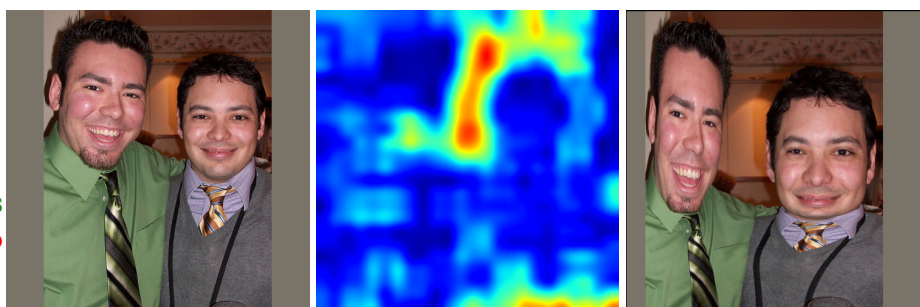


Figure 7: Additional qualitative examples on the POPE benchmark showing the input question, the original image, the processed perception map \mathcal{P} , and the perception-magnified image, arranged from left to right. The answers generated by greedy decoding and PM are compared, with correct answers highlighted in green and incorrect ones in red.

will need some attention mapping mechanism between visual tokens and patches. However, we can resolve this problem by introducing gradient backpropagation to address the implicit relationships between the image and tokens, refining the visual attentions and construct the perception map, similar to the method proposed in (Chefer et al., 2021). Overall, such limitation does not reduce the contribution of this work, as the idea of adaptive-magnification can also always be used in aligned with interpretability tools designed for different

VLMs. Alternatively, one could use PM in ways similar to (Yu et al., 2024), where any compatible model could be used as external guidance to the decoding-adaptive magnification.

K GPT-Assisted Evaluation

We adopt the GPT-assisted evaluation prompts directly from prior studies. For reference-based evaluation, we follow the prompt used in (Liu et al., 2023), with an example of the prompt and corresponding evaluation result provided in Table 8. For

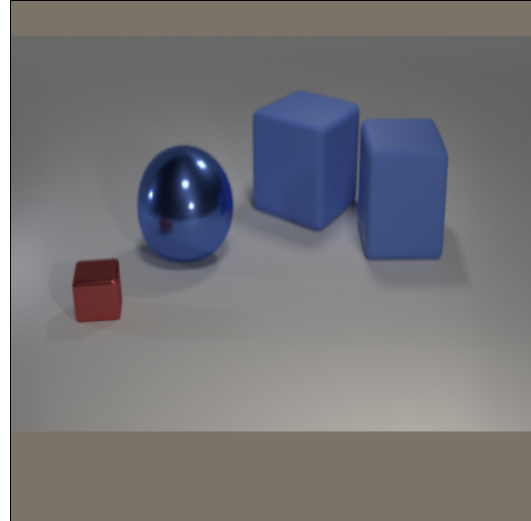
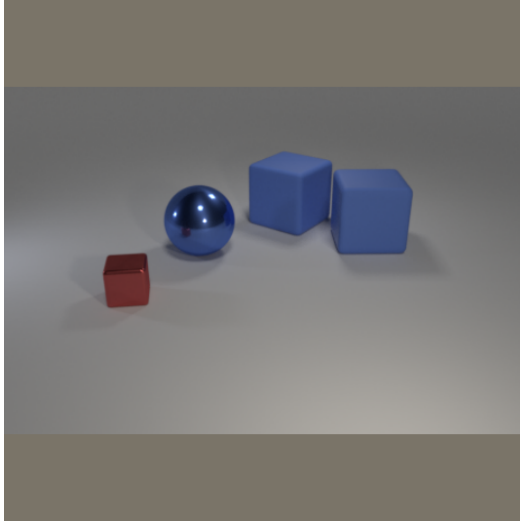


Figure 8: A sample CLEVR-style image (left) and its corresponding magnified visual input (right). For tasks involving shape or size-related attributes, the proposed PM may introduce unintended distortions due to unregulated magnification. In this example, magnifying both the red cube and the blue sphere results in an enlarged sphere and a deformed cube, which may negatively affect model predictions.

vision-based evaluation, we use the prompt from (Huang et al., 2024), with an example shown in Table 9.

Evaluation Protocol

Each tested method is evaluated under two settings:

- **Reference-based Evaluation:** GPT-4o is given the question along with a detailed textual description of the image.
- **Vision-based Evaluation:** GPT-4o is given the question and the original image.

In both settings, GPT-4o scores two answers: one from the tested method and another from GPT-4 (used as the reference). Each answer is rated independently on a scale of 0 to 10. The final evaluation score is computed as:

$$\text{Relative Score} = \frac{\text{Score}_{\text{method}}}{\text{Score}_{\text{GPT-4}}}.$$

This normalization ensures comparability across questions of varying difficulty.

Evaluation Criteria

The evaluation criteria differ slightly between the two settings:

- In **reference-based evaluation**, GPT-4o assesses overall content quality, coherence, and informativeness of the response relative to the image description.

- In **vision-based evaluation**, GPT-4o is explicitly instructed to focus on the *accuracy* of visual grounding and the *level of detail* present in the answer.

L LLM Usage Statement

We used GPT-4.5 exclusively during the writing phase of this paper, for the purpose of polishing the manuscript. Specifically, we provided prompts such as: “Revise the grammar and academic wording of this paragraph, and list aspects to be improved, including suggestions on how to improve them.” All outputs were subsequently reviewed and manually revised to ensure that technical accuracy was preserved. No part of the research ideation or experimental analysis relied on LLM assistance.

Prompt for Reference-based Evaluation

[Context]

An aerial view of Diamond Head in the Hawaiian Islands.

[Question]

What is the name of this famous sight in the photo?

[Assistant 1]

The famous sight in the photo is Diamond Head.

[End of Assistant 1]

[Assistant 2]

The famous sight in the photo is the Na Pali Coast, which is a scenic coastal area on the island of Kaua'i in Hawaii.

[End of Assistant 2]

[System]

We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. The user asks the question on observing an image. For your reference, the visual content in the image is represented with a few sentences describing the image.

Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.

Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space.

In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgement.

Response:

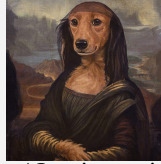
10 2

Assistant 1 correctly identified the famous sight in the photo as Diamond Head, which is consistent with the context provided. The response is accurate, relevant, and directly answers the question with the appropriate level of detail given the context.

Assistant 2, on the other hand, incorrectly identified the sight as the Na Pali Coast, which is not mentioned in the context. This response is inaccurate and irrelevant to the question, as it does not align with the description of the image provided. Additionally, the details about the Na Pali Coast, while informative, are not applicable to the context of Diamond Head. Therefore, Assistant 2's response is not helpful in this scenario.

Table 8: Prompt for reference-based evaluation.

Prompt for Vision-based Evaluation



You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:

1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.

2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details.

Please output the scores for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgement.

[Assistant 1]

The intended effect of this painting might be to create a humorous and whimsical twist on the famous Mona Lisa artwork by featuring a dog dressed in her attire, which can evoke feelings of amusement and curiosity.

[End of Assistant 1]

[Assistant 2]

The intended effect of this painting is to create a humorous and whimsical scene by combining a dog with a human-like appearance, such as a Renaissance-style painting. The dog is wearing a Renaissance-style dress and a hat, which adds a touch of humor and creativity to the image. The painting might be intended to evoke amusement and curiosity in the viewer, as it is not common to see a dog dressed in such a manner. The artist might have wanted to challenge the viewer's expectations and create a unique, memorable piece of art.

[End of Assistant 2]

Output format:

Accuracy: <Scores of the two answers>

Reason:

Detailedness: <Scores of the two answers>

Reason:

Response:

“9”: **Accuracy: 8 9**

*Reason:**

Assistant 1 accurately describes the image as a humorous twist on the Mona Lisa featuring a dog, but lacks specific details about the attire. Assistant 2 provides a more detailed and accurate description, mentioning the Renaissance-style dress and the humorous intent, which aligns well with the image content.

Detailedness: 6 9

*Reason:**

Assistant 1 gives a brief description focusing on the humorous aspect but lacks detail about the attire and style. Assistant 2 offers a richer description, noting the Renaissance-style dress and hat, and elaborates on the artistic intent, providing a more comprehensive understanding of the image.”

Table 9: Prompt for vision-based evaluation, note the image is processed by GPT-4o for evaluation.