# PatchPrune: Reducing Hallucinations in Vision Language Models by Pruning Redundant Image Patches

**Changyan Liu**[*] 2023190905023@STD.UESTC.EDU.CN
*Glasgow College, University of Electronic Science and Technology of China, Chengdu, China*

**Editors:** Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

## Abstract

Large language models (LLMs) have advanced significantly in natural language processing, and vision language models (VLMs) have extended this progress to tasks like image captioning and visual question answering (VQA). Despite this success, VLMs often generate hallucinated or factually inconsistent contents. Traditional methods focus on improving model reasoning by modifying the inference procedure, but we propose a new approach: PatchPrune, which dynamically prunes redundant or uninformative image patches, using a composite importance score based on activation magnitude and feature entropy. As shown in Figure By reducing input noise, PatchPrune enables the model to focus on relevant features, improving the accuracy and reliability of its outputs. Experimental results show that PatchPrune enhances multimodal reasoning and mitigates hallucinations effectively.

**Keywords:** Vision Language Models, Hallucinations, Multimodal Reasoning, Patch Pruning

## 1. Introduction

Large language models (LLMs) have achieved remarkable success across various tasks, propelled by the exponential scaling of training data, model parameters, and computational resources. Building upon these advances, vision-language models (VLMs) have emerged as a powerful extension that bridges visual and linguistic modalities. By jointly training on image-text pairs, VLMs such as CLIP, BLIP,Vision-LLMv2, mPLUG-Owl, and LLaVA (Ye et al., 2023; Liu et al., 2023; Ye et al., 2024) have demonstrated impressive performance in various multimodal tasks, including image captioning, visual question answering, and open-ended multimodal dialogue. This rapid progress has not only broadened the landscape of multimodal AI applications but has also brought new opportunities for cross-modal reasoning, understanding, and interaction.

However, despite these significant advancements, a pressing challenge remains: the issue of hallucination—where the model generates outputs that are factually incorrect, semantically inconsistent, or unsupported by the visual input. This limitation undermines the reliability and trustworthiness of VLMs, especially in high-stakes or real-world applications. In response, a growing body of research has been dedicated to understanding and mitigating hallucinations in VLMs.

Several recent works have proposed innovative methods to address this challenge. For instance, DoLA reduces hallucinations by contrasting the logits from higher and lower transformer layers during decoding, thereby amplifying factual signals captured in the upper layers without requiring external supervision or model fine-tuning. VCD (Leng et al., 2024) tackles object hallucinations by leveraging a contrastive approach between the model outputs given original and visually distorted inputs. This training-free method reduces reliance on spurious language priors and dataset biases. OPERA (Huang et al., 2024) introduces a novel decoding strategy that penalizes excessive reliance

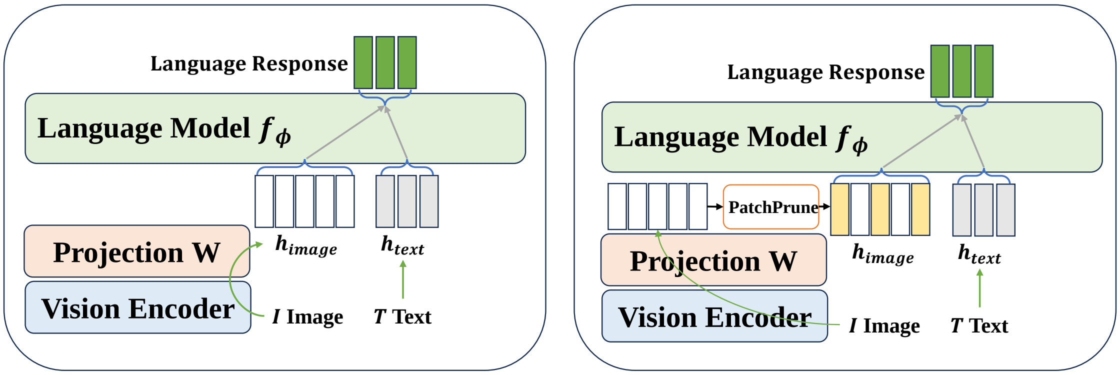

Figure 1: Comparison between the original LLaVA pipeline (left) and our proposed PatchPrune-enhanced pipeline (right). PatchPrune prunes less informative image patches (highlighted in yellow) and retains representative patches to improve multimodal reasoning.

on summary tokens during beam search and reallocates attention retrospectively to improve output fidelity—again without requiring additional data or retraining. DAMO (Wang et al.) takes a different route by introducing a momentum-based decoding mechanism, which accumulates activations across transformer layers to enhance visual grounding, thereby producing more faithful predictions across a variety of benchmarks.

While prior work has largely focused on modifying decoding strategies or enhancing reasoning capabilities at inference time, we take a different perspective: could the input itself—specifically, the abundance of image patches—contribute to hallucination? In many VLMs, images are divided into a large number of fixed-size patches and encoded indiscriminately, regardless of their individual informativeness. This can result in an input space saturated with redundant or irrelevant information, potentially overwhelming the model and impairing its ability to attend to truly salient regions. We hypothesize that this patch-level noise may play a significant role in hallucination.

To address this, we propose PatchPrune, a lightweight yet effective method that dynamically prunes uninformative or redundant image patches before multimodal fusion. As illustrated in Figure 1, PatchPrune selectively removes low-utility patches (highlighted in yellow) and retains only the most informative ones, leading to a more focused and robust representation. Our method leverages a composite importance scoring function that evaluates each patch's relevance by jointly considering its activation magnitude and feature entropy. Patches with low activation or minimal information content are pruned, thereby reducing the cognitive load on the model and improving its capacity for grounded, accurate reasoning.

Comprehensive experiments across multiple multimodal benchmarks validate the effectiveness of PatchPrune. Not only does our method reduce hallucination rates, but it also improves overall task performance, demonstrating better alignment between visual inputs and generated outputs.

In summary, our contributions are as follows:

- We introduce PatchPrune, a novel method for addressing hallucinations in VLMs by dynamically pruning uninformative or redundant image patches based on a composite importance scoring function.
- PatchPrune improves multimodal reasoning by reducing input noise, enabling the model to focus on the most informative features, leading to more accurate and reliable outputs.

## 2. Method

In this section, we first present the preliminary of the multimodal decoding process in vision language models (VLMs). Then, we formulate the patch pruning task as a patch selection problem based on feature importance. Finally, we introduce a composite importance scoring function that evaluates each patch's contribution using both activation magnitude and feature entropy.

### 2.1. Preliminary

Given an image $I$ and a text $T$, the image and text are first passed through separate encoders to extract their respective feature representations. Specifically:

$$\begin{cases} h_{\text{image}} = f_{\text{image}}(I), \\ h_{\text{text}} = f_{\text{text}}(T), \\ h_{\text{fusion}} = f_{\text{fusion}}(h_{\text{image}}, h_{\text{text}}) \end{cases} \tag{1}$$

Here, $f_{\text{image}}(\cdot)$ and $f_{\text{text}}(\cdot)$ are the visual and textual encoders, respectively, while $f_{\text{fusion}}(\cdot, \cdot)$ integrates the image and text features into a joint representation $h_{\text{fusion}}$.

The fused representation $h_{\text{fusion}}$ is used as the initial input to the transformer layers for multimodal reasoning. Let $h_t^j$ denote the hidden state at the $t$-th token and the $j$-th transformer layer. The inference proceeds as follows:

$$h_0^0 = h_{\text{fusion}}, \quad h_t^{j+1} = f_j(h_t^j), \quad j = 0, \ldots, N-1 \tag{2}$$

Where $f_j(\cdot)$ represents the operation of the $j$-th transformer layer, including self-attention and feed-forward components.

Finally, the output $o_{t+1}$ is generated by applying a language head $\varphi(\cdot)$ to the final hidden states $h^N$, which is used to predict the next token:

$$o_{t+1} = \text{softmax}(\varphi(h_t^N)) \tag{3}$$

The model follows an autoregressive decoding strategy, where the prediction of each token depends on the previous context.

### 2.2. Task Formulation

Given an image representation in the form of an embedding tensor $\mathbf{X} \in \mathbb{R}^{N \times D}$ (denoted as $h_{image}$ in 2.1), where $N$ denotes the number of spatial image patches, $D$ is the dimensionality of each patch embedding vector, and we denote the set of patch embeddings as

$$\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}, \quad \mathbf{x}_i \in \mathbb{R}^D \tag{4}$$

We aim to identify and remove a subset of uninformative or redundant patch tokens using a dynamic masking mechanism. Specifically, we define an importance scoring function:

$$\mathcal{I} : \mathbb{R}^D \to \mathbb{R} \tag{5}$$

which assigns a scalar score $s_i = \mathcal{I}(\mathbf{x}_i)$ to each patch, capturing its contribution to the global image representation. Based on the resulting scores $\{s_i\}_{i=1}^{N}$, we retain the top-$(1-r)N$ patches and discard the bottom $rN$ patches, where $r \in (0, 1)$ denotes the masking rate.

### 2.3. Importance Score Function

To effectively reduce input redundancy while preserving informative content, we introduce an importance scoring mechanism that evaluates the relevance of each image patch embedding $\mathbf{x}_i \in \mathbb{R}^D$. This score determines whether a patch should be retained or masked during subsequent reasoning. The importance score $s_i$ is computed by aggregating multiple signals that reflect different aspects of the patch's informativeness.

**Activation Magnitude (Energy-based)** The first component of the scoring function is based on the magnitude of the embedding vector. Intuitively, embeddings with low overall activation may carry less semantic or discriminative information. We compute this score as the $L_2$ norm:

$$\mathcal{I}_1(\mathbf{x}_i) = \|\mathbf{x}_i\|_2 \tag{6}$$

This energy-based signal provides a coarse yet effective heuristic for identifying patches that may have minimal influence on downstream tasks.

**Feature Entropy (Information Density)** To complement the energy-based score, we introduce an entropy-based metric that quantifies the internal diversity or uncertainty of the patch embedding. A patch with high entropy is assumed to encode more distributed and potentially informative features, while low-entropy vectors may indicate sparsity or peaked activations, signaling reduced information content.

We first normalize the embedding dimensions using a softmax transformation:

$$p_{ij} = \frac{\exp(x_{ij})}{\sum_{k=1}^{D} \exp(x_{ik})} \tag{7}$$

Then, we compute the Shannon entropy of the normalized distribution:

$$\mathcal{I}_2(\mathbf{x}_i) = -\sum_{j=1}^{D} p_{ij} \log(p_{ij} + \epsilon) \tag{8}$$

where $\epsilon$ is a small constant to avoid numerical instability. This entropy score captures the degree of uncertainty across the feature dimensions and serves as a proxy for information density.

**Composite Importance Score** To balance the contributions of both metrics, we define the final patch importance score as a weighted combination:

$$s_i = \alpha \cdot \mathcal{I}_1(\mathbf{x}_i) + \beta \cdot \mathcal{I}_2(\mathbf{x}_i) \tag{9}$$

Here, $\alpha$ and $\beta$ are tunable hyperparameters that control the relative influence of activation magnitude and entropy.

After computing the importance scores $\{s_i\}_{i=1}^{N}$, we retain only the top-$(1-r)N$ patches with the highest scores, where $r \in (0, 1)$ is the predefined masking rate. The remaining $rN$ patches are discarded, resulting in a reduced yet more informative set of visual tokens that are fed into subsequent reasoning modules.

## 3. Experiment

### 3.1. Experimental Setup

**Benchmarks** We adapt two benchmarks to evaluate our proposed method. MME is a benchmark designed to assess Vison Language Models (VLMs) on both their perception and cognition abilities across 14 subtasks, using manually curated instruction-answer pairs. For our evaluation, we primarily focus on the Perception category, which consists of 10 subtasks. POPE (Li et al., 2023) is an innovative pipeline created to evaluate object hallucinations. In this work, we utilize the official POPE benchmarks, which are generated from the MSCOCO and AOKVQA datasets.

**Models and Baselines** We adapt LLaVA 1.5 (7B) as our foundation to evaluate all benchmarks across various baselines. Our approach is compared against three baselines: Regular, DOLA, and VCD. Regular denotes the vanilla output of the VLMs. Visual Contrastive Decoding (VCD) (Leng et al., 2024) reduces object hallucination by contrasting the model's predictions given clean and intentionally distorted visual inputs. DOLA is adapted to the VLM setting by fixing the mature layer index at 32 and selecting from multiple premature layer indices (e.g., 0, 2, 4, 6, 8, 10, 12, 14), which is consistent with the original design.

**Hyperparameter Setting** In practice, we set $\alpha = \beta = 1$ to achieve a balanced trade-off. The masking rate $r$ is set to 0.05. Specifically, in the case of LLaVA 1.5 where $N = 576$, this corresponds to masking approximately $576 \times 0.05 = 29$ patches. Meanwhile, the temperature is set to 0 to perform greedy search for a fair comparison.

### 3.2. Results

**Analysis on MME Benchmark** As shown in Table 1, our approach outperforms other decoding methods. Compared to several widely used methods, DoLA only surpasses Regular decoding by 0.67 points, while VCD performs even worse than Regular by 31.32 points when the temperature is set to 0. In contrast, our method achieves the best performance, exceeding Regular decoding by 15.88 points, which demonstrates its effectiveness.

Table 1: Experiment results on MME bnchmark (Perception) across various decoding strategies using LLaVA 1.5, each column refers to a subtask, the best results are remarked with bold.

| Decoding | Existence | Count | Position | Color | Posters | Celebrity | Scene | Landmark | Artwork | OCR | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Regular | 190.00 | 160.00 | 138.33 | 165.00 | 140.48 | 135.88 | 156.25 | 161.50 | 118.50 | 125.00 | 1490.94 |
| DoLA | 190.00 | 158.33 | **143.33** | 165.00 | 139.46 | 133.24 | 157.00 | 161.50 | 118.75 | 125.00 | 1491.61 |
| VCD | 173.33 | 151.67 | 138.33 | 165.00 | 140.48 | **137.06** | 151.00 | 164.75 | **120.50** | 117.50 | 1459.62 |
| Ours | **190.00** | **165.00** | 138.33 | **170.00** | **140.48** | 121.76 | **156.25** | **166.00** | 119.00 | **140.00** | **1506.82** |

**Analysis on MM-Vet Benchmark** As shown in Table 2, our method also performs well on the more challenging MM-Vet benchmark. In terms of overall performance, VCD and DoLA achieve results comparable to Regular decoding. Specifically, VCD matches Regular, while DoLA performs 0.1 points worse. In contrast, our method surpasses Regular by 1.2 points. For individual subtasks, our method achieves performance comparable to other methods, except for the math task, where all

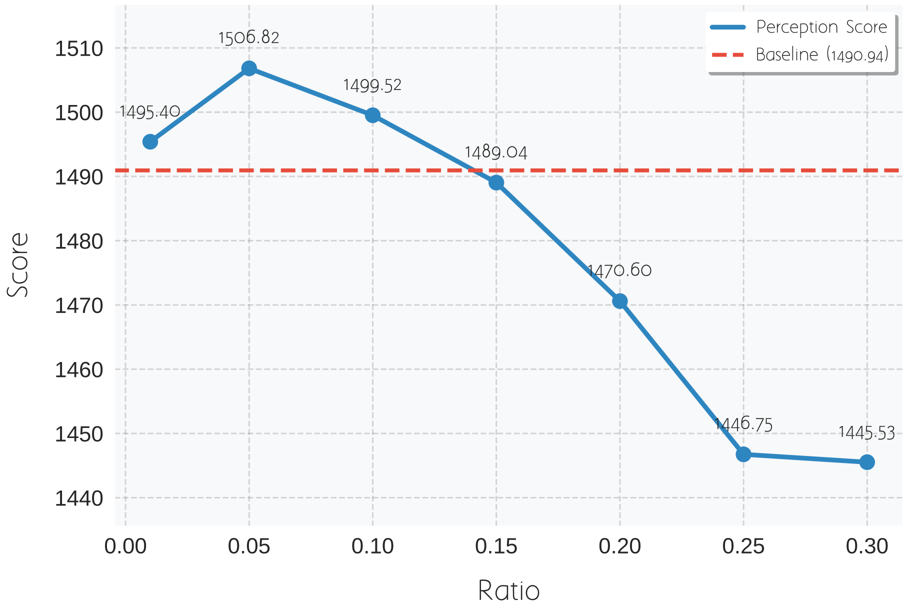

Figure 2: Ablation results on the MME benchmark evaluating the impact of different masking ratios $r$.

decoding strategies perform poorly. These experimental results on MM-Vet further demonstrate the effectiveness of our approach.

Table 2: Experiment results on MM-Vet benchmark across various decoding strategies using LLaVA 1.5.

| Model | Decoding | rec | ocr | know | gen | spat | math | total |
|---|---|---|---|---|---|---|---|---|
| LLaVA1.5 | Regular | 36.1 | 23.0 | 18.0 | 22.2 | 25.1 | **11.5** | 31.1 |
| | VCD | 36.1 | 22.4 | **21.0** | **23.1** | 28.4 | 3.8 | 31.1 |
| | DOLA | **36.5** | 22.3 | 18.1 | 23.0 | 25.7 | 7.7 | 30.8 |
| | Ours | 36.4 | **25.5** | 20.4 | 22.6 | **28.9** | 7.7 | **32.3** |

### 3.3. Ablation Studies

**Evaluation of Masking Ratio** $r$ To further evaluate the effect of different masking ratios $r$, we conducted experiments on the MME benchmark with values ranging from 0.01 to 0.30, while keeping other hyperparameters consistent. As shown in Figure 2, our method achieves the best performance when $r = 0.05$, demonstrating the reasonableness of our hyperparameter choice. Furthermore, when $r$ is set to other values, the performance of our method does not change significantly, indicating its robustness and generalization ability.

**Evaluation of $\mathcal{I}_1$ and $\mathcal{I}_2$** To evaluate the effectiveness of the proposed $\mathcal{I}_1$ and $\mathcal{I}_2$, we conducted separate experiments for each. By setting $\alpha = 0$, we first evaluate $\mathcal{I}_2$. With the optimal masking ratio $r = 0.05$, the result on the MME benchmark is 1501.53. In contrast, when setting $\beta = 0$, we evaluate the effectiveness of $\mathcal{I}_1$, yielding a result of 1499.08. Both results outperform the baseline, demonstrating the effectiveness of these two score metrics. Moreover, when combining these two metrics, the model achieves the best performance, further proving the superiority of the combined approach.

## 4. Conclusion

We proposed a novel method to address hallucinations in Vision Language Models (VLMs). Specifically, we designed a novel score metric to discard a fixed ratio of image embedding patches, retaining only the most reliable and important ones, thereby enabling inference without any additional training. The results on the MME and MM-Vet benchmarks demonstrate the effectiveness of our approach in addressing hallucinations in VLMs.

## References

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, pages 13418–13427, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pages 13872–13882, 2024.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NIPS*, 36: 34892–34916, 2023.

Kaishen Wang, Hengrui Gu, Meijun Gao, and Kaixiong Zhou. Damo: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. URL https://arxiv.org/abs/2408.04840.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.